Department of Functional and Comparative Genomics
Institute of Integrative Biology

# Optimising production of mannosylerythritol lipids by *Pseudozyma graminicola* through omic and molecular approaches

Stefany Solano González

Thesis submitted in accordance with the requirements of
The University of Liverpool for the degree of Doctor in Philosophy

November 2018


## ABSTRACT

Mannosylerythritol lipids (MEL) are biosurfactants produced primarily by basidiomycete fungi. MELs have potential applications in a wide variety of fields, ranging from medicine to industrial emulsifiers, making them an important target for investigation. MEL production is species-specific, with different proportions of the four variants MEL A-D, being produced and secreted. Of the MEL producers characterised to date all have a five-gene biosynthetic MEL cluster, directly involved on the synthesis of the molecule. However, the mechanisms involved in the regulation of these genes remains unclear. The work described in this thesis contributes to the understanding of MEL production by *Pseudozyma graminicola,* which primarily produces MEL-C. First, we sequenced and annotated the *P. graminicola* genome, from which we identified the MEL cluster. Based on comparative genomic analysis, we report *P. graminicola* as a potential biotrophic plant pathogen. Then, we developed a $^{1}$HMR based semi-quantitative method to confirm and monitor MEL production. Alongside this we identified the optimal conditions for the production of MELs using two fermentation systems. Transcription of the MEL cluster genes was monitored during MEL producing and non-producing fermentation conditions using RNAseq and RT PCR. We developed a transformation protocol, which was applied to deletion of *emt1* to validate the gene cluster function. We also used the same approach to investigate the roles of three transcription factors, *areA*, *gti1* and *pac2* with potential regulation roles in MEL production. Based on this analisis, deletion of *gti1* and *pac2* appear to result in reduced and enhanced MEL production, repectivly. Further detailed analysis of these mutant is now required to confirm and potentially exploit these findings. The developments described in this thesis contribute to better understanding the biology of *P. graminicola* associated to MEL production, providing the knowledge necessary for future development of high-yield strains.

# ACKNOWLEDGMENTS

In order to achieve a big task, there are important involved characters which play an important role in helping to accomplish the final outcome, therefore the following lines...

I would like to thank all the people who encouraged me at the beginning of this journey; my friends from home; Javi, Sierra, Don Carlos Alvarado. I also want to thank to all staff at the School of Life Sciences, UNA-CR. Special thanks to the PMI, Universidad Nacional, Costa Rica, for funding my research.

I would like to express my honest gratitude to my primary supervisor Prof Mark Caddick for guiding me in this research. Mark has been an outstanding supportive and patient supervisor throughout my PhD, he has giving me so much advice on improving work efficiency and scientific judgement. I am truly grateful for what I learned under his supervision. I would also like to thank my co-supervisor Alistair Darby for his scientific and bioinformatic guidance and Dr. Vasieva for her kind support at the beginning of my research. I express my gratitude to the team at CRODA, Doug Cossar, Phillipa Furnival and Anna Sobolewska for their intellectual contributions, material and collaboration. To both, my internal and external examiners for a wonderful viva and their kind contribution to the improvement of this document, Prof. Christiane Hertz-Fowler and Prof. Gary Jones.

I am deeply grateful to the CGR bioinformatics team, especially to Richard Gregory, Sam Heldenby, Mathew Gemmel and Luca Lenzi for the computational analysis advice and kind help on multiple occasions. Special thanks to the GeneMill staff, specially to Nichola Rockliffe and James Johnson, for the scientific input and molecular biology support and for the fun environment at the lab during the many hours I spent in there. I am also grateful to Marie Phelan for the training, guidance, advice and support for NMR technique; without her input the spectrometry analysis would not have been possible. Furthermore, I would like to extend my thanks to my colleagues who have become my friends Mark, Amber and the Darby group for making the many hours at work nicer and full of laughs. To Ewan for our beloved friendship. Louise, Gabi e Carina, obrigada do fundo do coração por estar sempre presente quando eu precisei. A amizade de vocês forem as pilhas que me mantinham em movimento. Gracias a Blanca por su apoyo incondicional y amistad.

Finalmente, gracias a mi familia: Mami, Papi, Roy y Sebas, por ser mi motor, mi refugio, mi fuerza y mi lugar seguro. Los amo. Su amor me permitió llegar al final de estas líneas, siempre les estaré agradecida.

*No te rindas que la vida es eso,*
*continuar el viaje,*
*perseguir tus sueños,*
*destrabar el tiempo,*
*correr los escombros y destapar el cielo*

M. Benedetti

*Don't give up, life is about that,*
*to continue the journey*
*to pursue your dreams*
*to unblock time,*
*to remove rubble and set off to an open sky*

M. Benedetti

# List of abbreviations

ACP: acyl-carrier protein
Cbx: carboxin
CPMG: Carr–Purcell–Meiboom–Gill
ddNTP: dideoxynucleotides
DE: differentially express
FA: Fatty acid
FAME: fatty acid methyl esters
FAS: fatty acid synthetases
FDR: false discovery rate
FR: Flanking region
Glu: Glucose
H: hours
HPLC: High-performance liquid chromatography
Hyg: higromycin
IIB: Institute of Integrative Biology
ITS: internal transcribed spacer regions
Lbs: pounds
List of abbreviations
MD: Multiple dose
MDS: Multidimensional scaling plot
MEL: Mannosylerythritol lipids
Mm: milimeter
MS: Mass spectrometry
NMR: nuclear magnetic resonance
PacBio: Pacific Biosciences
PCA: principal component analysis
Ppm: parts per million
QC: quality control
rRNA: ribosomal RNA
SD: single dose
SMART: Single Molecule Real Time
TAG: triacylglycerols
TCA: tricarboxylic acid cycle
TLC: Thin layer chromatography
WT: wild type
YM: yeast and mold agar.

TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

CHAPTER 1

# 1. FUNGI AS A SOURCE OF SECONDARY METABOLITES

The fungal kingdom is one of the largest and most diverse of eukaryotic kingdoms with an estimated of 1.5 to 5 million species (Choi and Kim 2017). Among the most common features associated with fungi, is the presence of a cell wall composed mainly of carbohydrate chitin (Bartnicki-Garcia, 1968). The production of ergosterol (Paterson 2005) and the synthesis of the amino acid lysine (Vogel 1964). However, the presence of these features can not be considered as definitive for the kingdom as for example the chitin synthesis pathway has been lost in several fungal groups (Bruns et al. 1992). In addition, the production of ergosterol has been found in protist that do not group phylogenetically with fungi (Thompson 1972) and the synthesis pathway for lysine, used as diagnostic character for fungi, is present in *Euglena*, a photosynthetic protist (Vogel 1964).

Fungal species display a wide variety of life cycles, metabolisms, morphogenesis (including hyphae, fruiting bodies, sexual and asexual spores) and ecologies. They are found in all temperatures, flora and are of great importance for the ecosystems through their functions of decomposing diverse substrates and of synthesising diverse classes of molecules (Petersen 2013).

Fungal species are classified in three main groups based on their proteomes, Monokarya, Ascomycota and Basidiomycota. The Monokaryotic group comprises Cryptomycota, Chytridomycota, and Zygomycota subgroups which do not appear to produce dikaryons during their life cycles. A dikaryon is the fusion of a pair of

compatible haploid nuclei of a fungus cell (Choi and Kim 2017). The Ascomycota group are dikaryon producers along with internal sexual spores called “asci” on top of fruiting bodies, members of this major group are Taphrinomycotina, Saccharomycotina, and Pezizomycotina. The basidiomycete group, also dikaryon producers, have their sexual spores formed externally on small pedestal fruiting bodies called basidia, and the subgroups Puccinomycotina, Ustilaginomycotina, and Agaricomycotina are part of this subgrouping (Choi and Kim 2017).

Basidiomycotas have an unicellular growth form called yeast which reproduces asexually by budding, fission or production of structures referred as ballistoconidia (Flegel, 1977; Fell et al., 2001). There are species which can alternate from yeast form to a filamentous growth form, known as hyphae (Steinberg, 2007). The life cycle of sexual basidiomycetes initiates with a haploid spore (basidiospore) that germinates to produce a free-living yeast which can reproduce asexually. In the presence of a compatible mating type, the yeast cells produce conjugation tubes which end up fusing to produce a dikaryotic hyphal cell (Figure 1-1).

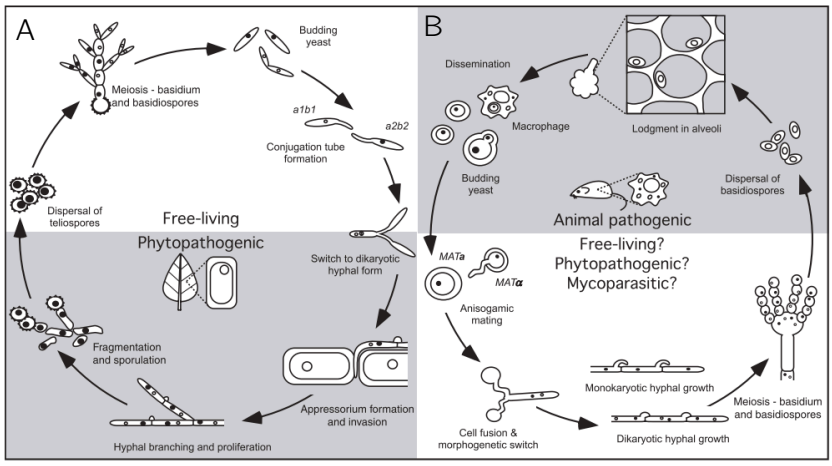



Figure 1-1. Life cycle diagram of basidiomycete. A) Free-living plant pathogenesis from *Ustilago maydis*. This panel shows the budding unicellular yeast forms which in presence of a pertinent partner fuses producing a dikaryon, then aprresorium (structure which will infect plants) will end up producing spores. B) *Cryptococcus neoformans* displaying a similar process as in for *U. maydis*. Essential for both fungi is the change from yeast to hyphal form, by sexual reproduction, to start the infection process. (Morrow and Fraser 2009).

Fungal capacity to synthesise a wide variety of molecules has been exploited since ancient times. The yeast *Saccharomyces cerevisiae* has been used worldwide in the production of fermented beverages, the domestication of this species is considered a pivotal event in human history (Johnson 2013). The development of science gave birth to fungal biotechnology, using the vast applicability of this eukaryotic kingdom. As an example, the Ascomycota group harbours *Aspergillus* species, predominantly used for food fermentation and production of enzymes, organic acids and bioactive compounds. *A. oryzae* is mainly used for the fermentation of soybeans, rice, grains and potatoes and the species *A. niger* is industrially used in the production of organic acids, including 99% of citric acid production worldwide (Van Bogaert et al. 2013; Johnson 2013; Magnuson and Lasure 2004). On the other hand, the basidiomycota group has also something to offer to fungal biotechnology. Among the beneficial attributes applied to industry are the production of carotenoids and fragrances, formation of important enzymes in pharmaceutical production and biotransformation, degradation of pollutants and bioremediation activity (Johnson 2013). As an industrial example, CRODA a UK based company, is involved in the production of coatings and polymers, lubricants and polymer additives, by the production of fungal metabolites by basidiomycete strains from the Basidiomycete family.

This low molecular bioactive metabolite produced by fungal strains, termed secondary metabolites, are not considered essential for the viability of their producers (Hoffmeister & Keller, 2007; Teichman *et al.* 2011) but facilitate nutrient transport, microbe-host interaction or act as biocide agent. Additionally, they have important roles in transcription, development and intracellular communication.

Therefore, understanding fungal strains is imperative in order to get access to these valuable metabolites. Although, our current knowledge of this group is very

limited to less than 7% on the plant-associated microfungi species (Mueller and Schmit 2007) and, despite the good characterisation of fungal model organisms such as *Aspergillus nidulans*, *Ustilago maydis*, *Schizosaccharomyces pombe* and *Neurospora crassa* (Hedges 2002), there is still a big gap of information (Hedges 2002). Nevertheless, the advances in science and its different applications have filled this gap of information by implementing omic techniques, by the means of recovering, characterising and interpreting the information on their genomes and transcriptomes. This has facilitated not only a better and detailed understanding of the fungal species in question by explaining their biology in their reproductive niches but has also allowed for better exploitation of their biosynthetic potential (Chen-Shan *et al*. 2013) through the identification of genes involved in the production of key secondary metabolites, facilitating the manipulation of the organism to increase yield or alter the final product.

## 1.1 Fungal Secondary metabolites

The study of fungal secondary metabolites began in 1922 with the characterisation of more than 200 mould metabolites, led by Harold Raistrick (Raistrick 1950). Nevertheless, it was until the discovery and development of penicillin , a secondary metabolite, that more efforts were focused on this topic (Keller, Turner, and Bennett 2005). These compounds have been proved to be of utility in a wide variety of fields, ranging from pharmaceuticals, as antibiotics: penicillin and cephalosporin produced by bacteria or immunosuppressants such as cyclosporines produced by *Tolypocladium inflatum*, through to environmental and industrial applications such as bioremediation and *biosurfactants*. (Abdel-Mawgoud *et al*. 2011, Brakhage 2013).

Natural surfactants, secondary metabolites known as biosurfactants, are surface-active compounds (SACs), produced by a wide range of organisms (Paraszkiewicz and Dlugoński, 2003; Holmberg, 2001). Biosurfactants were originally discovered as extracellular amphiphilic compounds produced during bacterial fermentation (Kitamoto *et al.* 2009, Soberón-Chávez and Maier, 2011). An amphiphilic molecule has a hydrophobic domain (usually fatty acids which can be saturated, unsaturated, linear, branched or hydroxylated) and an hydrophilic domain which is often a carbohydrate or peptide (Isoda *et at*, 1997; Irudayaraj *et al.* 2008).

There are five classes of biosurfactants classified mainly on the basis of their chemical structure and origin, such as: glycolipids, fatty acids, lipopeptides, polymeric and particulate. The key property of these compounds is that they lower surface and/or interfacial tension allowing the partition of water/oil or oil/air (Arutchelvi et al. 2008; Banat et al. 2010; Jezierska et al. 2018). In nature it is not well understood why these molecules are produced, nevertheless it is hypothesised that organisms will produced them when surface or interfacial tension changes are needed at the cell surface or the local environment such as erection of fruiting bodies, swarming of cells, gliding motility or for cell development (Jezierska et al. 2018).

An important feature of biosurfactants is that they have low critical micelle concentration (CMC), which is the concentration at which a surfactant aggregates into micelles (Figure 1-2). This CMC is usually lower than their chemical counterparts which makes them very efficient at small concentrations (Arutchelvi et al. 2008). Due to these features and their biodegradability, mild production conditions, and functional diversity, biosurfactants have attracted considerable commercial interest in recent years (Medrzycka and Karpenko 2009). Their ability to increase solubility of hydrocarbons is exploited commercially to stabilise or destabilise emulsions (Surekha *et al.* 2010), such as creams, ointments, pastes or balms in the health care industry (Troy et al. 2006). Compared to chemical

surfactants (carboxylates, sulphates and esters), Biosurfactants present higher structural diversity and lower toxicity. Importantly, they often retain specific activity under extreme conditions (pH, temperature and ionic strength) and can be produced at similar yields as chemical surfactants (Arutchelvi et al. 2008). Additionally, as in nature biosurfactants are often involved in complex social responses that control cell development, they also have important biological properties such as antitumor, antimicrobial and cell-differentiation activities (Rodrigues *et al*. 2006). For example, it has been demonstrated that specific biosurfactants direct human promyelocytic leukemia cells (HL60) to differentiate into granulocytes instead of promoting their proliferation. All these features make biosurfactants highly desirable within industry (Desai & Banat, 1997), their diverse properties providing a strong commercial and scientific reason to justify investment of time studying their properties and the underlying molecular mechanisms involved in their production.

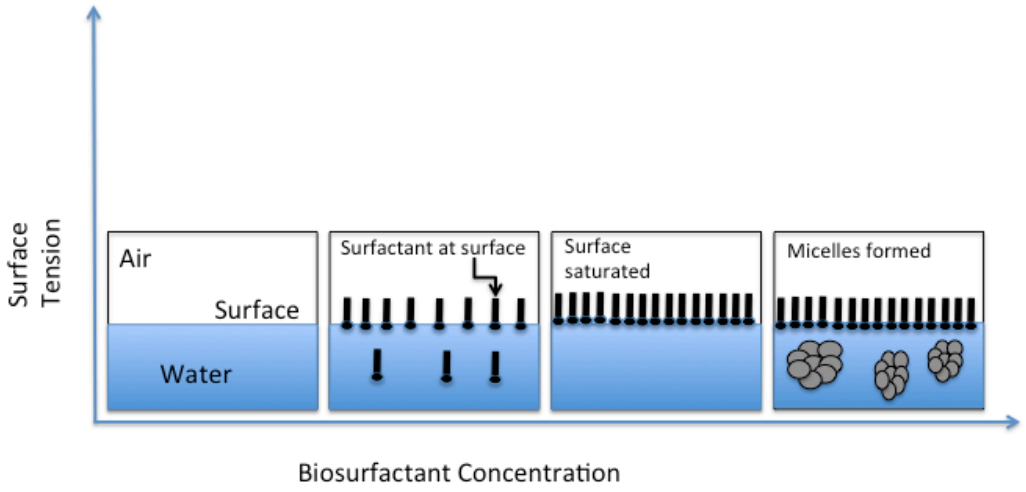



Figure 1-2. Cartoon representing critical micelle concentration concept. First slide showing air and water interface. Second slide showing biosurfactant molecules, third slide showing saturation of interface by biosurfactant molecules. Fourth slide showing micelle formation.

## 1.2 Mannosylerythritol lipids as biosurfactants produced by yeasts

Biosurfactants production has been described primarily in relation to bacteria, among the reported species *Pseudomonas* sp., *Acinetobacter* sp., *Bacillus* sp. and *Arthrobacter* sp predominate (Banat et al. 2010).

Among the different types of biosurfactants, the glycolipids have been most extensively studied because of their relatively high levels of production using renewable resources and their versatile biochemical properties (Morita et al. 2009). From this group, the Mannosylerythritol lipids (MELs) have gained particular interest over the past couple of years (Kitamoto 2008). All MEL molecules have the same sugar moiety mannosyl-erythitol, but they differ with respect to their fatty acid (FA), which can vary in terms of chain length and the level of saturation (Irudayaraj *et al.* 2008). Depending on the organism, MEL structure may vary with respect to the number and position of acetyl groups on the mannose, erythritol or both, amongst other features. The degree of acetylation has been used to classify the different forms of MEL produced: MEL-A (diacetylated), B and C (monoacetylated at the C4 and C6 positions, respectively) and MEL-D (completely deacetylated) (Irudayaraj *et al.* 2008) (Figure 1-3).

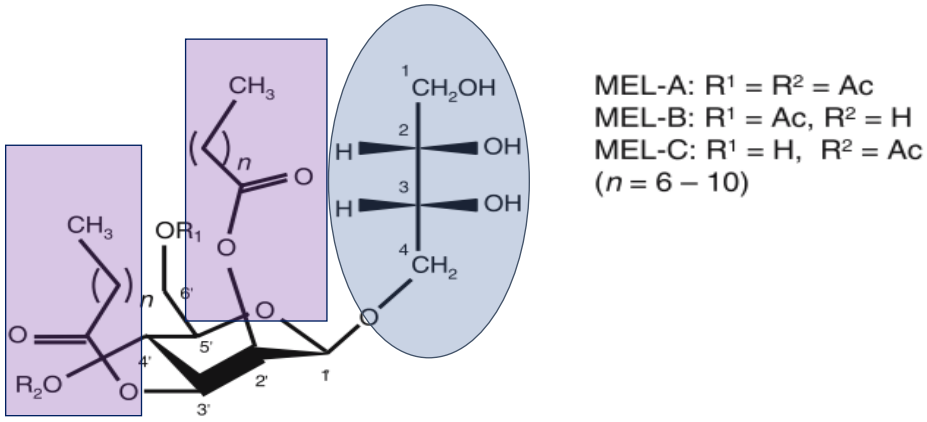



Figure 1-3. Chemical structure of MELs (Morita et al. 2006). FA marked with purple squares, sugars with blue circles.

Yeast strains of the genus *Pseudozyma* and *Ustilago* produce MELs abundantly (Kitamoto et al.1993, 2002; Hewald et al. 2006, Morita et al. 2007, 2009a; Teichmann et al. 2007; Arutchelvi et al. 2008; Bölker et al. 2008, Konishi et al. 2013, Faria et al. 2014, Yoshida et al. 2014, Saika et al. 2016). These basidiomycetes produce the glycolipid primarily when the culture reaches stationary phase, as it is not reported to be growth associated (Hewald et al. 2005, 2006).

MEL producing yeast make a mixture of the four types of the molecule and the proportion varies according to the species. For example, species that produce predominantly MEL-A are *P. aphidis*, *P. antactica* and *P. rugulosa*. MEL-B is mainly produced by *P. tsukubaensis*. MEL-C is the predominantly form produced by both *P. hubiensis* (65% of its final composition) and *P. graminicola* (85% of its final composition). It has been proposed that MELs function as an energy storage material in yeast (Dai Kitamoto, Isoda, and Nakahara 2002) or demonstrated for *U. maydis,* to enhance the availability of hydrophobic nutrients during the interaction with host (Hewald et al. 2005). However, the intrinsic biosynthetic regulation and natural role of MEL secretion remains vague (Günther et al. 2015).

## 1.3 Fatty acid production by microorganisms

Fatty acids are a major component of the MEL molecule. Due to the restricted knowledge on MEL biosynthetic regulation, understanding putative pathways involved in the production of precursors assembling the molecule are a good way approach to optimising its production.

The natural production of fatty acids (FAs) from alkanes has been reported in plenty of microorganisms (Dai Kitamoto, et al 1998) by three main metabolic pathways: 1) "*de novo* synthesis pathway" followed by β-oxidation, 2) "chain-elongation" pathway and 3) "intact incorporation pathway". The first requires an

increase in carbon source (glucose) and a starvation of nitrogen to take place. Under these conditions the production of FA is mainly due to a malfunction of the enzyme isocitrate dehydrogenase as a result of a blockage of the tricarboxylic acid (TCA) cycle which ends up as an increase in the production of Acetyl-CoA and oxaloacetate (Figure 1-4). The former is metabolised to produce triacylglycerols (TAGs), which are transferred to an acyl-carrier protein (ACP) to produce precursors on which fatty acid synthetases (FAS) act on, ending in aliphatic molecules of 16 carbons chain length (Athenaki et al. 2018). The second involves acetyl CoA carboxylase and malonyl CoA, which are used by FAS to elongate the FA chains by two carbons each round in the cycle and the third one as it name suggests does not increase the number of carbons from the chain regardless the substrate (Figure 1-4).

The intact elongation pathway, having similar principle as the previous explain pathway, couldn't explain the elongation of the FA chain (Kitamotoet al 1990, Athenaki et al., 2018) (Figure 1-4).

Nevertheless, none of these three pathways could justify the production of MELs. In order to elucidate this, an experiment using cerulenin (a strong inhibitor to *de novo* fatty acid synthesis) showed the production of MELs remained intact whereas when 2-bromooctanoic acid (a strong inhibitor of the fatty acid β-oxidation pathway) was used, the production of MEL was inhibited and the degree of this inhibition was directly proportional to the chain length of the supplied substrate (Yanagishita, Haraya, and Kitamoto 1998). Intriguingly, MEL producers accumulate tryacilglycerols (TAGs) intracellulary regardless of the carbon source used as other oleaginous yeast, which followed the *de novo* FA synthesis. In addition, when fatty acid methyl esters (FAMEs) with shorter chains are used as substrates, the TAGs yielded had $C_{14}$ to $C_{16}$ chain length, probably synthetized *via "de novo* synthesis pathway". If FAs with longer carbon chains (longer than FAMEs) as methyl pentadecanoate ($C_{16}$) were used the TAGs yielded were odd-chained length FAs, mainly synthetized *via* a "chain-elongation pathway" together

with “intact elongation pathway” for the TAGs formation ( Kitamoto et al. 1990.). These data suggested that more than one of the three FA biosynthetic pathways participated in the formation of TAGs and the pathway selection depends on the chain length of the substrate ( Kitamoto et al. 1990).

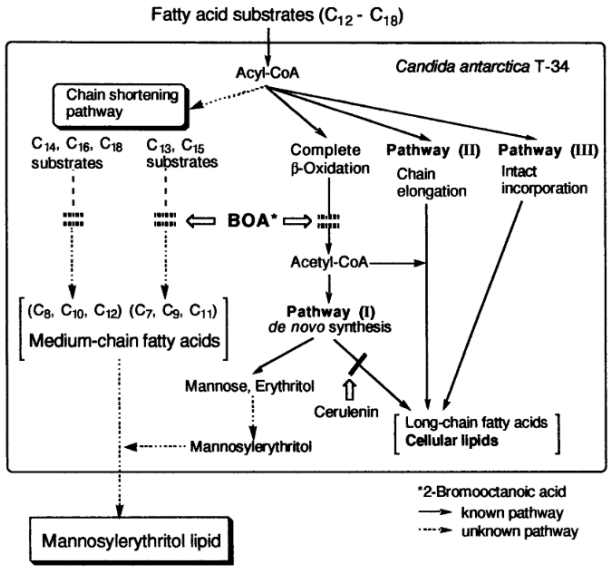
In the same study were TAGs pathways were analysed in *Candida antarctica* (changed to *P. antarctica*), they found the production of “unusual FA”: MELs ( Kitamoto et al. 1990). Kitamoto and Yanagishita (1993) could not be explained by the “elongation pathway” or by the “intact incorporation pathway” as they observed a decrease of $C_2$ regardless the substrate used. Likewise, the unchanged odd number carbon chain and unsaturation of the FAMEs showed the biosynthesis of MELs could not be due to “*de novo synthesis pathway*”. Therefore the proposal of a new fourth pathway in microorganisms was required to explain this glycolipid synthesis. In this respect, the “chain shortening pathway”, which partially oxidase FAs and shortens the carbon chain of their intermediate, was proposed as best candidate to explain the MEL biosynthetic pathway ( Kitamoto et al. 1990).

## 1.4 MEL biosynthetic cluster

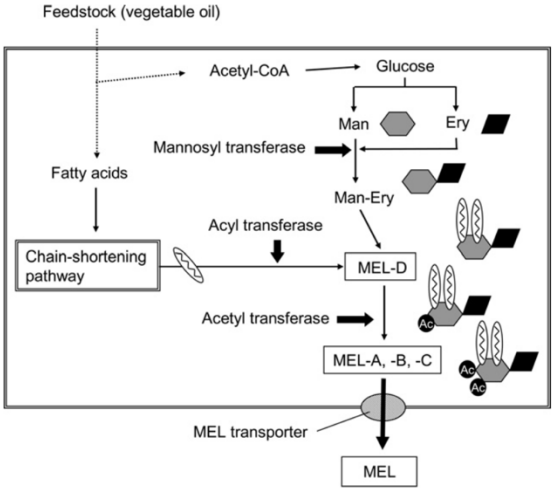
The biosynthetic cluster responsible for MEL production was first identified by Hewald and colleagues in *U. maydis* genome (2006). They accomplished this by implementing an expression analysis under nitrogen starvation using DNA microarray technology. They identified upregulated genes and the cluster was found due to the adjacent location of the genes to *emt1*, which was previously characterised as a glycosyltransferase (Hewald et al. 2005). This gave an insight of potential co-regulation, characteristic of clusters for secondary metabolite production. Subsequently, once the genes from the cluster were identified, deficient mutants were constructed to validate the function of the newly identified genes. By analysing the production of the mutants, they confirmed the function of the genes from the cluster (Hewald 2006). These genes encode a

glycosyltransferase (*EMT1*), two acyltransferases (*MAC1* and *MAC2*), an acetyltransferase (*MAT1*) and a transporter gene (*MMF1*) (Morita *et al.* 2014, Teichmann *et al.* 2007, Hewald *et al.* 2006, Morita *et al.* 2013, Konishi *et al.* 2013, Lorenz *et al.* 2014). The glycosyltransferase, which is essential for MEL biosynthesis, produces mannosylerythritol by connecting GDP mannose to erythritol (Morita *et al.* 2009, Morita *et al.* 2014). This sugar is then acylated with fatty acids by the acyltransferases at positions C-2 and C-3 (Figure 1-3), process required for the production of MEL. By constructing deletion mutants was demonstrated that both acyltransferases are essential for MEL biosynthesis and deletions of either *mac* genes resulted in a complete loss of the lipid in *U. maydis* (Hewald et al. 2006). However, it remains unclear which enzyme acylates which carbon, as well as their order of activity. Depending on the MEL type produced *MAT1* adds acetyl groups prior the extracellular secretion of the molecule/molecules by the transporter MMF1 (Hewald et al. 2005, 2006). However, many processes related to the metabolic function and gene expression of the cluster remain unclear (Günther *et al.* 2015*)* (Figure 1-4).

A




B




C

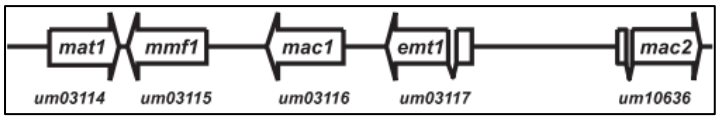



Figure 1-4. Fatty acid biosynthetic pathways in microorganisms and MEL cluster. A) Condense version of fatty acid production pathways by *P. antarctica* T-34. B) Presumptive synthesis path way of MELs by *P. antarctica.* C) MEL biosynthethic cluster (Kitamoto et al 1990; Morita et al. 2009).

Homologous gene clusters have been found in a number of other basidiomycetes. Phylogenetic analysis of these MEL clusters has shown that the genes involved are highly conserved among the *Pseudozyma* genus and related species (Morita et al. 2013).

Nevertheless, even though MEL production and regulation are not well understood, many studies have focussed on the physical parameters affecting production,

among those the carbon source being fatty acid and/or glucose as nitrogen are key nutrients as the genes from the cluster were upregulated under nitrogen starvation.

### 1.4.1 Fatty Acids as carbon source

Among the factors that affect MEL production, carbon source is an important one. *Pseudozyma* strains can use almost any vegetable oil, with soybean and olive oil being mainly used. It has been shown that fatty alcohols or acids with chain-length of $C_{12}$ to $C_{16}$ as substrates yielded MELs with a FA chains lengths of $C_2$, $C_4$ and $C_6$. Interestingly even-numbered FA substrates ($C_{12,}$ $C_{14}$, $C_{16,}$ or $C_{18,}$) resulted in MEL chain lengths of $C_8$, $C_{10}$ and $C_{12}$, whereas odd-numbered FA produced acids of $C_{7,}$ $C_9$ and $C_{11}$ (Kitamoto et al. 1993). The profiles of carbon chains tend to be the similar between different species, if the type of MEL produced (i.e MEL-A, MEL-B, MEL-C, MEL-D) is the same (Arutchelvi et al. 2008; Morita et al. 2009) whereas the sugar moiety remains mainly unchanged. This suggests that biosynthetic pathway producing the sugar moiety is more conserved than the hydrophobic moiety (Morita et al. 2006).

### 1.4.2 Sugars as secondary carbon source

The type of sugar used as substrate also affects the yield and chain length of the MEL molecule. The use of glucose is generally preferred by the organisms, resulting in a higher yield when compare to other substrates, such as xylose (Faria et al. 2014). This is probably due to the simplicity of directly utilising glucose rather than having to breaking down xylose for its further assimilation. Although for species such as *P. rugulosa* using erythritol as second carbon source increased the MEL yield, when compare to glucose alone (Morita et al. 2006). In most *Pseudozyma* species when glucose is available in the fermentation medium the cells utilise this preferentially, carbon catabolite repression ensuring optimal use of

carbohydrates for energy. However, FAs are also required to form the MEL backbone. Consequently, uptake of both substrates is generally required during the fermentation in order to produce MELs.

Many studies report the use of edible vegetable oils, such as soybean, for MEL production (Adamczak and Bednarski 2000; Isoda et al. 1997; Morita et al. 2006, 2013, 2014; Yanagishita, Haraya, and Kitamoto 1998; Yoshida et al. 2014). However, from an industrial point of view this is uneconomic due to the high price of this FA sources. For this reason, optimisation using waste feedstock is a priority for commercially viable MEL production.

### 1.4.3 Nitrogen source

In *U. maydis* nitrogen starvation conditions aid to identify the cluster. Additionally, has been shown that when cells are nitrogen starved neutral lipids accumulate as discrete deposits (Guo et al. 2009; Liu et al. 2013), named lipid bodies (LB) as a response to a change in the nutrient supply, allowing the cell to alter the production of LB to meet the energy demand and entering into autophagy (Zavala-Moreno et al. 2014). Reports in the literature indicates that lipid accumulation in oily yeasts has an ideal C/N ratio of 100/1 as an excess of glucose may be channelled into TAG synthesis and this accumulation, as mentioned before, due to nitrogen starvation triggers TAG accumulation in the cell. This leads to an increase of citrate in the cytosol, which is subsequently converted to pyruvate, AcetylCoA and subsequently to oxaloacetate in the TCA cycle in the mitochondria. The presence of these molecules serve as precursors for lipid synthesis in microorganisms, such as *de novo* synthesis pathway and the chain-shortening pathway (Zavala-Moreno et al. 2014).

In order to get MELs produced the carbohydrate concentration is usually kept high whereas nitrogen is required to be depleted. A concentration of 0.2% of $NaNO_3$ has been shown to provide the highest yield for *P. antarctica,* whereas nitrogen sources, which lead to more extreme acidification of the media (e.g $NH_4Cl$ or

$(NH_4)_2\,SO_4$), resulted in low MEL yields (D Kitamoto et al. 1990) probably due to a dropped of the pH (Kitamoto et al. 1990b; Rau et al. 2005a, b).

## 1.5 Chemical approaches to study MEL molecules

### 1.5.1 Recovery and Isolation

As mentioned before MELs are produced as a mixture of its four types and this ratio is species-specific. *U. maydis* produces mainly MEL-A and another type of surfactant named cellobiose lipids (Hewald et al. 2006) whereas *P. rugulosa* NBRC 10877 produces from soybean oil a mixture of MEL A (68%), MEL-B (12%), and MEL-C (20%) (Morita et al. 2006). In addition, the carbon chain might differ in length, meaning the recovery of biosurfactants involve a heterogeneous mixture of different molecules with an intrinsic amphiphilic nature. This requires the use of solvents (such as n-hexane and methanol) to remove oil and fatty acids and later on chloroform to separate the organic phases. Then recover the phase of interest using for example a silica gel (Niu et al. 2017). Another method to get a clean fraction is to implement a preparative high-performance liquid chromatography (HPLC) equipped with silica gel columns. However this procedure incurs in a substantial loss of the product (Udo Rau et al. 2005).

### 1.5.2 Identification and characterisation

Due to a lack of commercially available MEL standards, the identification and characterisation of MELs has been achieved by implementing mainly three analytical techniques: thin layer chromatography (TLC), nuclear magnetic resonance (NMR) and mass spectrometry (MS).

### 1.5.3 Thin layer chromatography (TLC)

TLC is a simple and low cost technique that allows the detection of most of the components in a sample. In general terms requires a solvent system which will make a sample move through a silica plate and will separate the compounds from the sample based on their interaction and hydrophobicity (Costanzo 1997). The most frequent technique to identify MELs is by the anthrone method, which involves a mixture of sulphuric acid and anthrone reagent which stain the sugar moiety of MELs, therefore MEL molecules with different acetylations will migrate differently through the silica plate (Dai Kitamoto et al. 1990). Usually HPLC is coupled to this technique by eluting the desired compound from the plate.

### 1.5.4 Mass Spectrometry

In this analytical technique a previous ionisation of the biological samples is required to allow the spectrometer to monitor the trajectory of its atoms and molecules in an electric or magnetic field (Lössl, van de Waterbeemd, and Heck 2016). This provides a relative intensity of the measured compounds, which translates into peak signals for each metabolite. As a result of the ionisation process each compound from each molecule will generate a fingerprint of the original molecule by producing different peak patterns. Currently in the market a wide range of both instrumental and technical variants are available, differing mainly in the ionisation and mass selection methods (Alonso et al 2015). In respect to MEL analysis, this technique allows the quantification and identification of different variants of carbons present in the hydrophobic chain of the MEL molecule. Fan and collaborators (2014) analysed the products from *Pseudozyma aphidis* ZJUDM34 and identified the fatty acid profile quantifying the relative content of each carbon. They found this species produces MEL molecules with a carbon chain ranging from eight up to 20 carbons with a variance of saturated and unsaturated forms (Fan et al. 2014). Hence this technique has a very good

power of resolution and provides very detailed information of the analysed compounds. However, its drawbacks are an important matter too. Among these are: the necessity of good and pure standards to compare to, pure samples to work with, an adequate tuning up of the machine to detect the m/z ratio beyond the noise signal, a good database to match the obtained hits and the high cost of the equipment and required expertise to manipulate it (Alonso et al 2015; Lössl, van de Waterbeemd, and Heck 2016). All these reasons make the setting of MS as a daily routine not too appealing.

### 1.5.5 NMR

NMR analysis is an analytical technique based on the spin nature all nuclei possess. In this respect the most abundant nucleus and therefore, most common element conforming molecules is hydrogen, being the most well studied isotope ($^{1}H$) abundantly present in nature (Marion 2013). The targeted sample (cells, fluids, media) brought under a magnetic field (ranging in the market at fields of 200, 300, 400, 500, 600, 800 up to 950 MHz) will spin according to the charge on its nuclei, and this spin will be observed by the detector coupled to the magnet (Marion 2013). The recording of this shifts in junction to thermodynamics tables of chemical shits for determined metabolites (under determine circumstances) are constantly reported. The sample preparation varies according to the aim, for instance if intracellular metabolites are the target a prior extraction is required, whereas if extracellular metabolites are analysed this procedure could be omitted. The usual downstream analysis after the post-instrument processing of the spectrum involves a 1) quality control in which offsets spectrum are removed, 2) calculation of intensity values (by data point on each peak or over segmented regions – called bins -), 3) conversion of spectral data to analytical measurements, usually arranged in a table containing observation/samples (in rows) and variables/frequencies (in columns), 4) normalisation of the data, to adjust as required the spectral intensities, 5) scaling of the data and 6) the statistical

analysis to model the data (Craig et al. 2006). A predefined pattern of signals can be fitted to the recorded spectrum in order to calculate the intensity values. Under this scheme, for MEL analysis plenty of work has been done, characterising the molecule by one dimension ($H^1$). Nevertheless, as with MS this technique also has its limitations, especially when the aim is to identify different MEL species (A, B, C, D) as MEL-B and MEL-C have the same arrangement; not being distinguishable by one dimension. In the other hand the set up involves less steps than MS.

Implementation of analytical techniques such as NMR and MS has aid on this identification, unravelling a high productivity variation between species from the same group or even within the same species, according to the conditions. A good approach to accomplish the understanding of MEL production, despite the high variability, discussed in previous sections, is to employ a strain which ratio of production is as homogeneous as possible and focusing on its genetics and potential production dynamics.

## 1.6 The study of *Pseudozyma graminicola* as a vehicle to investigate MELs

*Pseudozyma graminicola*, a yeast belonging to the division *Basidiomycota* (Golubev *et al*. 2007) was isolated for the first time from grass in Moscow, Russia (Golubev, Sugita, and Golubev 2007). This species has the potential to produce valuable metabolites, including MELs. Its main product, which constitutes 85% of the final mixture, is MEL-C; showing the highest hydrophilicity among the MELs. This feature makes it highly advantageous for downstream applications such as water-in-oil emulsifiers or washing detergents (Morita et al. 2008). Prior to this work *P. graminicola* had been poorly characterized and the genome was unsequenced. Therefore, its sequencing, assembly, and annotation are required in order to get suitable information allowing its further characterisation, modification and engineering.

### 1.5.6 Study of Fungal genomes

The sequencing of fungal genomes allows the understanding of the diversity, behaviour, regulation and expression of genes encoding enzymes and pathways involved in the production of relevant compounds. Sequencing of the first fungal species took place in 1996 with the model species *S. ceverisae* (Mohanta and Bae 2015).

Fungal genomes are relatively small (<100 Mb) compared to the genome size of animals and plants (Tunlid and Talbot, 2002). The fungal genome size varies according to the group they belong to, with the Ascomycota being the smallest (~36.91 Mb) and the Oomycota the biggest (74.85 Mb) (Mohanta and Bae 2015).

The constant efforts on sequencing fungal genomes fall into two main groups, one with medical relevance (Cuomo and Birren 2010) and the other with industrial relevance (Grigoriev 2011). The filamentous group of fungi has gain more attention as it comprises species involved in the production of compounds such as antibiotics, organic acids and industrially relevant enzymes, used in food, detergents and other industries highly applicable to biotechnology industry (Tsang 2014). The first sequenced filamentous fungi genome was *Phanerochaete chrysosporium* in 2004 (Martinez et al. 2004). This led to sequencing of more genomes from the *basidiomycota* family, were MEL producers belong to.

Therefore, the increasing on sequencing, assembling and annotation of fungal genomes, from the *basidiomycota* group is necessary as many more ascomycetous yeast genomes have been studied. Consequently, more studies focused on basidiomycete fungal genomes have led to a greater understanding of their biology and aids on the development of yeast for industrial purposes, such as MEL production, leading researches to decipher metabolic pathways allowing its further manipulation (Johnson 2013).

However, development of predictive tuned-up tools able to handle the intrinsic nature of fungal genomes are required; due to features such as tight gene spacing promoting overlap of untranslated adjacent genes, presence of multiple short introns. Unlike animal genomes, there is a poor understanding of signal transcription start and stop, and translation start sites in fungal genomes. Furthermore, comparative analysis of fungal genomes shown fungi are very divergent (Galagan et al 2005) even between members of the same genus, at the genomic level (Mohanta and Bae 2015). In addition to this divergence, events of genome duplication and translocation have had a major impact in the evolution of yeast, demonstrating how dynamic fungal genomes are on nature (Dunham et al. 2002; Koszul 2004).

This justifies the need to increase the number of high quality available genomes from the basidiomycete group, as to date only 391 genomes are deposited at the NCBI from the approximate 35 000 reported (Choi and Kim 2017; de Vries et al. 2017) and from these only eight species correspond to MEL producers. Hence applying a pipeline which allows the sequence, assemble and annotation of a strain like *P. graminicola* adds value to the development of industrial and medical biotechnology.

## 1.7 Genome Sequencing

DNA sequencing started with the well-known methods developed by Sanger and colleagues (Sanger *et al.* 1977) and the method developed by Maxam & Gilbert (1977). The principle of this system consisted in using dideoxynucleotides (ddNTPs) that block the DNA extension; when ran in parallel with individual ddNTPs the outcome will function as an autoradiography to infer the original nucleotide from the template. The main drawbacks of this technique consisted of very low automatisation and short sequencing length.

### 1.7.1 Short-read Sequencing

Technological advances led to a 2nd generation of sequencing or short-read sequencing, which was mainly divided into two technologies: 1) sequencing by ligation and 2) sequencing by synthesis. In the former the main flow comprises a probe sequence bound to a fluorophore that hybridizes to a DNA fragment and is ligated to an adjacent oligo-nucleotide for imaging. The spectrum emitted by the fluorophore reveals the identity of the bases complementary to its specific position within the probe.

The sequencing by synthesis involved the use of the luciferase protein, releasing pyrophosphatase proportionally to the amount of DNA sequenced, therefore named pyrosequencing (Nyrén & Lundin 1985). In general terms a polymerase is used and a signal, being usually a fluorophore or a change in ionic concentration, identifies the incorporation of a nucleotide into an elongating strand by emitting light (Goodwin, McPherson, and McCombie 2016). In both approaches the DNA is amplified multiple times on a solid surface aiming for a strong enough signal distinguishable from background noise. The technologies available for sequencing by synthesis had variants on the market such as Roche/454, Torrent and Illumina (Goodwin, McPherson, and McCombie 2016). Whereas sequencing by ligation was more restricted to SOLiD technologies. Therefore the high-demand for the Illumina sequencing platform indulged a dramatic cost reduction and allowed an increase in automation, throughput and sampling depth (Cui *et al*. 2010).

The main drawback with these technologies was the relatively short length of reads, which usually did not exceed 800 base pairs (Gupta 2008, Goodwin *et al*. 2016). This is an important limitation for genome assembly (discuss in further sections), particularly limiting the ability to span long repetitive elements, making the assembly less contiguous. For this reason, long reads are key in assembling contiguous sequences with few gaps (Eid *et al*. 2009, Lee *et al*. 2014).

### 1.7.2 Long-reads Sequencing

The constant improvement of sequencing market evolved to developed 3rd generation sequencing; however, the threshold defining second and third is not entirely clear. Although we will consider third generation as Heather and Chain (2016), as "the technologies capable of sequencing single molecules, negating the requirement for DNA amplification shared by all previous technologies". Under this assumption, single molecule-sequencing technologies, of which the Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) platform is currently the most used.

SMRT is capable of producing long reads of up to 50 kilo bases (kb) with an average read length over 10 kb with an accuracy of 87% as demonstrated on the rice genome (Lee et al., 2014, Koren & Phillipy, 2015). SMRT sequencing involves a single DNA polymerase molecule attached to the bottom of a zero-mode waveguide (ZMW), each of which corresponds to one of many thousands of cavities in a thick metal film of a ZMW (Thudi *et al*. 2012). The DNA molecules to be sequenced are utilised by the DNA polymerase molecules as templates for DNA synthesis. During the DNA synthesis, each of the four nucleotides is identified by a different labelled fluorophore that is attached to the phosphate group. Excitation of the fluorophore is achieved by laser-beam-mediated illumination, allowing its identification. The unincorporated nucleotides remain in a section that is not illuminated and therefore do not light up. After the incorporation of each nucleotide the phosphate-dye complex is released and diffused out of the detection area; this enables an elongation of thousands of nucleotides within minutes. As this process occurs simultaneously across all the ZMW the sequence of thousands of DNA molecules can be determined simultaneously in real time (Gupta 2008) (Figure 1-5).

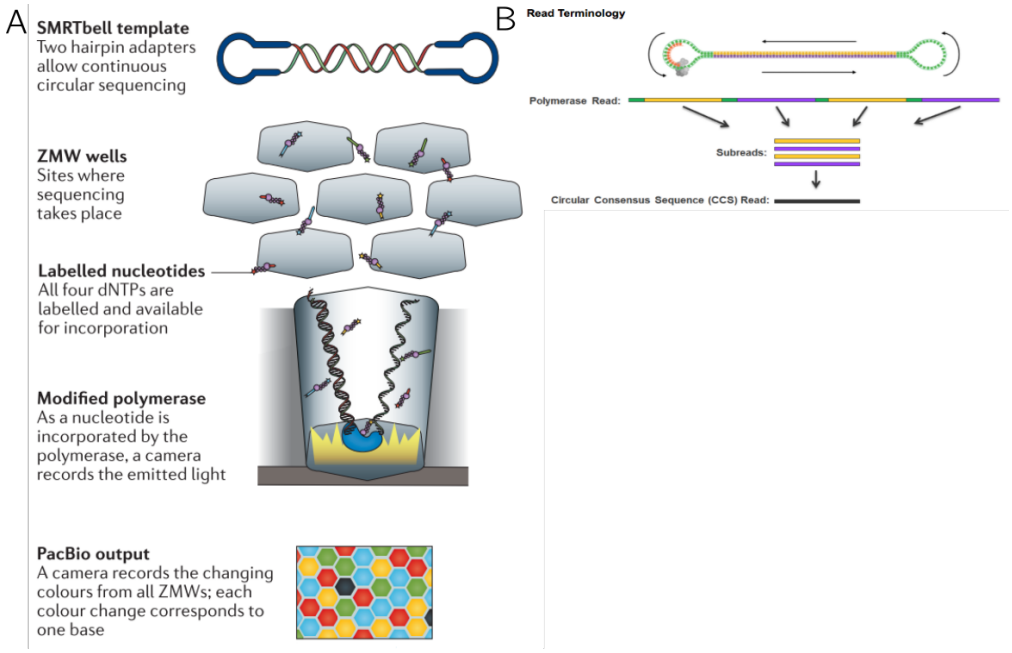



Figure 1-5. Single molecular real time (SMRT) sequencing from PacBio. A) Hairpins allowed double and circularise amplification, multiple hairpins are attached to ZMW wells. B) Representation of multiple subreads obtained by implementing this circular amplification (Goodwin, McPherson, and McCombie 2016).

This technology delivers extended reads, which allow the resolution of large structural features. In addition, can span complex and repetitive regions, eliminating the ambiguity in the positions or size of genomic elements. In transcriptomic studies (discuss in further sections), these long reads can span entire mRNA transcripts letting researches identify the precise connectivity of exons and discern gene isoforms (Goodwin et al 2016). Nevertheless, there is another option on the market: Oxford Nanopore Technologies (ONT), which offers a small mobile phone-sized USB device capable to sequence viruses, such as Ebola (Hayden 2015), additionally to plenty of applications. However, reports of poor quality are yet an inconvenient (Heather and Chain 2016).

## 1.8 Genome Assembly

After the sequencing is done, the genomic assembly is the next step. As mentioned previously, *P. graminicola* has not been previously sequenced, hence we took the

*de novo* approach, defined as the bioinformatic process of assembling sequenced fragments into a full contiguous genome for an organism that has not been sequenced previously. This process involves powerful computer algorithms to piece together the reads into continuous stretches of sequence known as contigs. Another important concept is scaffold, defined as longer stretches of joined contigs (Ekblom and Wolf 2014). In principle the outcome of the process is to align the long reads end to end by matching up overlapping areas allowing the extension of the assembly by including all overlapping sequences (Sohn and Nam 2016).

In order to account for possible errors high sequence coverage is key. Sequence coverage consists on the average number of times a region has been "covered" or aligned by independent reads. The higher this is the more likely it is that overlaps of unique sequences will occur between reads allowing each position in the genome to be linked up (Ekblom & Wolf 2014). The longer the sequence reads, the longer the overlaps expected and therefor the probability of these overlaps being informative (i.e. unique) increases. To optimally utilise these data, a long-read consensus algorithm has been developed by PacBio; the Hierarchical Genome-Assembly Process (HGAP) (Chin et al, 2013). This uses the longest reads as "seed reads" to which all other reads are mapped. Then a preassembly is performed, which converts the seed reads into highly accurate preassembled reads, by correcting errors using the shorter overlapping reads from the same library. Finally, the consensus step takes place, where all initial data is used refining the assembly to produce the genome assembly (Chen-Shan et al. 2013).

Once the genome has been assembled the final step is to place and orient the contigs or scaffolds onto putative chromosomes. When there is no reference genome for comparison, the best approach is to use orthology and gene order from related species. This must be done with care as chromosomal rearrangements may have occurred, even between very closely related species (Ekblom & Wolf 2014).

### 1.8.1 Gene Calling as structural annotation

Gene calling is usually a challenge, referring to the process of identifying the regions of genomic DNA comprising a gene; especially for eukaryotic genomes due to their large size and the presence of introns and transcript heterogeneity (Cantarel *et al*. 2008). For this reason, various pipelines have developed to optimise this process. Usually the gene-calling process is divided into two distinct phases. The first one comprises a "computational phase" were any possible evidence is aligned to the genome and *ab initio* and/or evidence-driven gene predictions are generated (Yandel & Ence 2012). In non-model organisms this phase is usually based on exons or transcripts; this process is automated using aligners such as Bowtie and Tophat in virtue of a lack of gene models (Yandell and Ence 2012). The second step is the "annotation phase", were all information is then combined following the rules determined by the implemented annotation pipeline to provide gene predictions (Ekblom & Wolf 2014).

The chosen technology, depends on the biological question to answer. For example among the most used platforms for RNA seq experiments PacBio and HiSeq are currently the golden standard. Nevertheless, PacBio has a higher error rate compared to HiSeq, which limits the *novo* transcript identification. Instead HiSeq provides more abundant (higher depth) and accurate reads (Conesa *et al*. 2016), a key feature for the detection of low abundance transcripts.

On the other hand, the incredible long reads produced with PacBio (exceeding 10 kb in length) are useful for *de novo* genome assemblies (Schadt et al. 2010; van Dijk et al. 2014). This feature also allows the discrimination between methylated and un-methylated bases, as the polymerase pauses for longer when attempts to elongate DNA containing modified bases, this pause is detected by a metric called "interpulse duration" (Flusberg et al. 2010).

In consequence the implementation of both techniques (long and short sequencing reads) gives a very good resolution to elucidate new transcript models and gene quantification.

## 1.9 Transcriptome Analysis

The term *transcriptomics* was used for the first time in 1990s and is defined as the group of techniques to study an organism's transcriptome (the sum of all its RNA transcripts). This process captures a real time snapshot of the total transcripts present in a cell under a particular circumstance (*i.e* time point, condition) (Lowe et al. 2017).

An important factor delaying the analysis and interpretation of transcriptome (RNA-seq) data is the numerous genes that have no function assigned (even for well-studied model organisms). This becomes critical for genomes whose annotation is mainly based in gene predictions and propagative methods (David et al. 2008), usually the case when using only genomic data.

Another important element to consider is the presence of multiple transcripts and isoforms for genes, which expression and regulation is shaped by internal or external stimuli (Conesa et al. 2016). Therefore measuring the expression of an organism's gene in different tissues, conditions or time points provides vital information on how genes are regulated and gives detail of an organism's biology. From an annotation point of view, this is key to deducing the function of previously unannotated genes (Lowe et al. 2017) and to creating new transcript models (Robertson et al. 2010), enhancing the quality of a prior *de novo* assembly.

An essential step for RNA-seq analysis is to get the targeted RNA; being mRNA the object for RNA sequencing experiments. In order to separate these molecules from the ribosomal RNA (rRNA), which is the most abundant species, either the polyadenylated tails of the mRNA are captured or the rRNA is removed directly (Conesa et al. 2016). In the first case only mRNAs, microRNAs, and snoRNAs

transcripts are captured using oligo(dT) primers or beads as the polyadenylated tail will hybridise to the complementary dT sequence. Whereas the rRNA depletion works the opposite way, selecting out the rRNA with "specific oligos" and then binding this oligos to beads for their removal (Zhao et al. 2014).

Once isolated, the mRNA is reverse transcribed in order to get a complementary DNA molecule (cDNA). A variety of commercial kits are compatible with the Illumina platform; currently the most commonly used for RNA (or cDNA) sequencing. The Ultra Directional RNA Library prep kit (NEB) includes a step to obtained cDNA by priming random primers to the isolated mRNA and with the addition of buffer and a retro transcriptase the first cDNA strand is generated. Then this first strand will be used as template to generate the second strand. This kit allows the creation of paired-ends, meaning the sequence is obtained in both 5'-3' and 3'-5' directions, derived from the first and second strand. Following this, the next step involves cDNA fragmentation, as the available sequencing technology requires of short sequence length (up to 500 base pairs -bp-). The fragments are linked to adapters and amplified by PCR. Then the molecules are size selected within a range of 150 bp to 300 bp long (Mortazavi et al. 2008).

Libraries sequenced from both ends (paired-end reads) provide much more information than single-end reads (Trapnell *et al.* 2012) as the former ones improve the mappability of the reads hence the transcript identification.

Once the libraries are sequenced, the next step is to align the RNA reads contiguously. Illumina technology creates reads in a length range of 50 to 300 bp long. Optimally, each particular transcript will be represented by a large number of small reads. Provided there is sufficient depth (coverage) of sequence, these reads are likely to overlap, allowing accurate alignment and assembly of the full transcript sequence. If a reference genome sequence is available the reads can be mapped onto this. This gives a better output due to the ability of programs such as Tophat to align reads across splice junctions (Yandell *et al.* 2012).

Furthermore, after the reads are aligned, the next step is to assemble them into transcripts, the most commonly used tool for such task is Cufflinks. This software implement algorithms for assembly and expression quantitation that has more accuracy when dealing with paired-end reads than with single-end reads (Trapnell *et al.* 2012). This occurs due to an increase in the likelihood of Cufflinks assembling correctly a full transcriptome by having information from both ends.

## 1.10 Functional Genome Annotation

In addition to gene-calling another important task is the functional annotation of genomes, which facilitates comparative analysis and is essential to the biological interpretation of sequenced data. Due to the massive amount of data generated these analyses are relatively time consuming and represent a major bottleneck (Jang-il & Jin-Wu 2016). Consequently different tools and pipelines are constantly developed in order to push forward the annotation.

The best-known functional annotation schema for gene and protein sequences is Gene Ontology (GO) (Ashburner et al. 2000). The main goal of this consortium lead initiative was "to produce a structured and precisely defined common controlled vocabulary for describing the roles of genes and products in any organism" (Ashburner et al. 2000).

Blast2Go is a bioinformatic tool for DNA or RNA sequences, facilitating automated functional annotation based on GO vocabulary (Götz et al. 2008). This makes use of other available tools such as Blast, InterProScan (Jones et al. 2014) and KEGG pathways. The principle is to integrate all these tools in order to transfer functional equivalency of orthologous genes to the genome being annotated. This is required, as the vast majority of the genes identified by sequencing will never be experimentally validated (Koonin 2005). Validity of this approach is dependent on the assumption that true orthologs are likely to retain identical function over evolutionary time, and there is good experimental evidence

in support of this (Li et al. 2003). An important factor to consider is the choice of database used to retrieve the information, as this will significantly affect the quality of information used and reliability of the annotation produced. Usually the best option is to use primary databases such as SwissProt, which are restricted to highly curated proteins (Bairoch et al. 2004). The use of Blast2Go requires of a licence, which means of an extra cost for some laboratories, resulting inconvenient some times.

Therefore the implementation of bioinformatic tools such as InterproScan, when using primary databases, enhances the quality of the annotation procedure making it accessible without the need to purchase any licence. As this is a building process, were the broader the database the broader the scope of annotation, which is very important to consider during the annotation process.

## 1.11 Thesis Objective

To use genomic and transcriptomic approaches to identify the biosynthetic genes for MEL-C production in *P. graminicola*. Primary among these are the MEL biosynthetic genes, which are likely to occur in a gene cluster, the function of which will subsequently be validated. This will also offer the potential for the application of directed mutagenesis to develop high yielding strains. The goal is to clarify and fully characterize the molecular biology underlying MEL biosynthesis and develop approaches for the identification of similar gene clusters in novel unannotated organisms or metagenomes.

### 1.11.1 Aims of Thesis

Due to the high value and applicability of MELs; especially MEL-C being the most hydrophilic from the four available forms we aimed to investigate *Pseudozyma graminicola,* a MEL-C producer. To achieve this, we aimed to:

A. Sequence and assemble the genome of *P. graminicola*. The functional and structural annotation of *P. graminicola* strain allowed us to identify the MEL cluster genes; along with providing valuable information for establishing key molecular genetic approaches that can be used to improve the quality of annotation of other genomes and the exploitation of MEL production by this yeast. Additionally, by functional and comparative genomics to other model fungi species is possible to get a better understanding of its yet unknown biology.

B. Obtain and utilise transcriptomic data to inform gene calling and identification of the MEL-C biosynthetic pathway. Our goal with this was to monitor the genes from the MEL cluster during MEL producing and non-producing conditions over four days of fermentation to understand the mechanisms behind its production and regulation.

C. Confirm the function of the putative MEL-C biosynthesis pathway by directed gene deletion and monitor expression of these and related genes involved in precursor production. We implemented $^{1}$H NMR to semi-quantified the MEL production by deficient strains and reported the general observations about changes in morphology for these mutants, which to the best of our knowledge have not been previously attempted in a *Pseudozyma graminicola* strain.

D. To undertake strain development to increase MEL-C production. We implemented *U. maydis* deletion cassettes and homologous recombination technique to knock out relevant genes to MEL biosynthesis.

# 2. GENOMIC SEQUENCING AND ANNOTATION OF *PSEUDOZYMA GRAMINICOLA*

## 2.1 INTRODUCTION

*P. graminicola* is a major MEL-C producer. This feature is highly advantageous for downstream applications, as the product is more homogeneous, the recovery of the molecule is expected to be less complicated. MEL-C can be used as water-in-oil type emulsifiers or washing detergents (Morita et al. 2008). Prior to this work *P. graminicola* had been poorly characterized and the genome was unsequenced. Most characterised members of the *Ustilaginomycotyna* group are also MEL produces, with *Sporisorium scitamineum* being the one known exception. The genome sequences and assemblies are publicly available for a number of these species: *Ustilago maydis* (GCA_000328475.2), *Pseudozyma antarctica* strain JCM10317 (GCA_000747765.1), *Pseudozyma antarctica* (GCA_000334475.1), *Pseudozyma aphidis* (GCA_000517465.1), *Sporisorium scitamineum* (GCA_001010845.1), *Pseudozyma hubeiensis* SY62 (GCA_000403515.1) (Konishi, Hatada, and Horiuchi 2013a, 2013b; Lorenz et al. 2014; Morita et al. 2013, 2014; Saika 2014; Taniguti et al. 2015).

The sequencing of these smut fungi has increased in the past five years, although most of these remain as draft genomes, meaning their assemblies are comprised of a large number of contigs. The quality of these assemblies is low when compared to that of the model organism, *U. maydis*. These sequences were obtained mainly using the HiSeq platform, which allows deep coverage but is dependent on relatively short reads. Other sequencing platforms, such as PacBio, can produce much longer reads, up to and exceeding 10 kb in length, which are useful for de *novo* genome (Koren and Phillippy 2015).

*P. graminicola,* is a non-model organism, therefore there are no pre-existing gene models for this species, to which a comparison can be made. For this reason, a good approach to implement is the *ab initio* automated gene prediction in which mathematical models, rather than external evidence, are used to identify and determine the genomic structure of genes (Yandell and Ence 2012). Additionally, other pipelines utilise RNA-seq data as evidence to improve the gene prediction. As an example, Braker, a recently reported highly accurate pipeline for gene prediction (Hoff *et al.* 2015) for unsupervised RNA-Seq-based genome annotation implements tools such as GeneMark-ET (Lomsadze et al. 2014) and AUGUSTUS (Stanke et al. 2008). In this pipeline GeneMark-ET performs an iterative unsupervised training using RNA-Seq reads to subsequently generate the *ab initio* gene predictions. From this a subset is used as "evidence" to train AUGUSTUS (Stanke et al. 2008). Then, this software incorporates information derived from the mapped RNA-Seq reads into the prediction step.

## 2.2 Chapter aims

I. To sequence and assemble *Pseudozyma gramicola* genome using the PacBio platform, in order to get high quality long reads.
II. To get transcriptome sequencing data using the HiSeq platform and mapped it back to *P. graminicola* sequenced genome.
III. To identify reference (s) species to which *P. graminicola*, can be aligned in order to rearrange contigs.
IV. To perform both gene-calling and functional annotation for the assembled genome.

### 2.2.1 Chapter description

This chapter describes the genome sequencing of *Pseudozyma gramicola* using the PacBio platform, which provided us with high quality long reads. We then assembled the genome using, as references, two closely related fungi. We undertook transcriptome sequencing, using the HiSeq platform and mapped these data back to the genome sequence. Finally, we perform both the gene-calling and functional annotation for the assembled genome. For the gene-calling, we used the Braker pipeline, as we followed the *de novo* approach. In order to obtain the biological annotation for the genome we implemented the Blast2Go suite and extended these results by running bioinformatics tools such as InterProScan, OrthoMCL, and KEGG orthology in parallel.

## 2.3 MATERIALS & METHODS

### 2.3.1 Culture conditions

An individual colony of *P. graminicola* strain CBS10092 was inoculated into 50 mL of growth media (Table 3-10) at 30 °C and 200 rpm in an orbital incubator (120 rpm) for 48 hours at the IIB facilities. To harvest the cells, the culture was centrifuged at 3000x g for 15 minutes at room temperature. The pellet was frozen on liquid nitrogen and stored at -20$^{o}$ C for later extraction.

### 2.3.2 Genomic sequencing and assembly

#### 2.3.2.1 DNA extraction

For extraction of genomic DNA 1g of cells (wet weight) were ground in a mortar and pestle with liquid nitrogen. The powder was transferred to 1 ml of extraction buffer (Tris 100mM, NaCl 1.4M, EDTA 10mM, CTAB 2%, pH 8.0) in a 2 ml eppendorf tube. The mixture was incubated for 10 minutes at 65 ° C, and then

centrifuged for 2 min. The supernatant was transferred to a fresh tube containing 4 µL of RNAse (100mg/ml), mixed gently and incubated at 37˚C for one hour. A total of 700 µL of chloroform:isoamyl alcohol (25:1) was added and centrifuged after vortexing, the upper phase (~500 µL) was transfer to a new tube. This last step was repeated twice. The upper phase was transfer to a new tube and an equal volume of isopropanol added. The mixture was invert a few times and left on ice for 5 min and then centrifuged for 5 min. The liquid was removed and the pellet was washed with 500 µL of 70% ethanol, vortexed for 1 sec and then centrifuged for 2 min at maximum speed. The remaining liquid was removed again without disrupting the pellet and the tube left on the bench to air-dry. After all the ethanol had evaporated the DNA was re-suspended in 200 µL of sterile $H_2O$.

To assess the concentration and quality of the genomic DNA, it was run overnight on an agarose gel (0.5%, w/v) at 30 V, 400 mA, in TAE 1X. As the molecular weight marker the 1 kb extension ladder (Invitrogen) was used.

#### 2.3.2.2 PacBio sequencing

*P. graminicola's* genomic libraries were made for the Pacific Biosciences (PacBio) platform (Mosher *et al.* 2014; Koren & Phillippy, 2015) using four SMRT (single-molecule, real-time) cells with P6.C4 chemistry to a 50x coverage. This allows the sequencing of DNA fragments of up to 10kb (Koren and Phillippy 2015). This sequencing was carried out by the Centre of Genomic Research (CGR), Liverpool.

#### 2.3.2.3 Genome assembly

The PacBio HGAP assembler version 3 was implemented and carried out by the CGR, Liverpool. The rearrangement of *P. graminicola* contigs was conducted using progressive MAUVE version 2.4.0 (Darling et al. 2004) using the assembled genomes of *Ustilago maydis* strain 521 (GenBank ID: 70) and *Sporisorium reilianum* (GenBank ID: 10890) as references. We then checked for any

improvements on *P. graminicola's* contig arrangement utilising the NUCmer program, MUMmer 3.0, with default settings (Delcher et at. 2003). The dot plots for NUCmer were generated by mummerplot (Kurtz et al. 2004) using the layout option (-l) to run the arrangement of scaffolds.

### 2.3.3 RNA sequencing

#### 2.3.3.1 Fermenter growth conditions and sampling

This part of the experiment was performed at CRODA facilities. After 48 hours of growth at 30 °C and 120 rpm (see section 2.3), approximately 50 ml of *P. graminicola* seed culture was inoculated into two 5 L fermenters (Applikon) containing 2 L of producing media (Table 2-2). The only difference between the two fermenters is that one also included fatty acid (FA) (80g/L) for the induction of MEL production. During incubation one fermenter was feed with glucose (non-induced) and one with both FA and glucose (induced). The sugar feeds were at the rate of 2g/hr/3L using a 50% glucose solution filter sterilised. The FA feed was at 2g/hr/3L. The fermentation was run for 117 hours and samples of 50 ml were taken from each fermenter at 24, 48, 72, 96 and 117 hours. The fermentations were conducted twice, independently, but due to the failure of an air compressor in one of the fermenters the first experiment stopped after 72 hours. This means we got, for each condition, duplicate samples for 24, 48, 72 hours and a single sample for 96 and 117 hours.

#### 2.3.3.2 RNA extraction

Cells were harvested by centrifugation, washed twice with PBS (to remove excess of FA from the media) and immediately frozen with liquid nitrogen. RNA extraction utilised the RNeasy Midi Kit (Qiagen Hilden, Germany), according to the manufacture's protocol. The frozen pellet was ground in a mortar and pestle with liquid nitrogen. The powder was resuspended in 4 ml of lysis buffer RLT

which was supplemented with 0.01%v/v of β-mercaptoethanol. As genomic DNA depletion is already part of this kit no extra DNAse treatment was perform. The concentration of the RNA was measure using the Qubit HS RNA assay (Thermo Scientific) and the integrity was analysed with the Agilent 2100 Bioanalyzer using the RNA pico kit (Agilent, Waldbronn, Germany).

#### 2.3.3.3 Poly(A) selection for mRNA enrichment

The RNA depletion of the samples was perform using the NEBNext Poly(A) mRNA magnetic isolation module (New England BioLabs, Ipswich, MA, USA) according to the manufacturers' protocol. We used 5 µg of total RNA. The concentration of the mRNA was measure using the Qubit HS RNA assay (Thermo Scientific) and the integrity was analysed with the Agilent 2100 Bioanalyzer using the RNA pico kit (Agilent, Waldbronn, Germany). Only samples that reached a ribosomal integrity number (RIN) higher than 8 were used for the next step.

#### 2.3.3.4 cDNA synthesis and library preparation

The NEBNext Ultra Directional kit (New England BioLabs, Ipswich, MA, USA) was used for the cDNA synthesis with 1 to 0.6 ng of mRNA as input material. A total of 17 cycles of PCR amplification was conducted using multiplex primers (indexed primer used to give a unique identifier to each library). The concentration of the cDNA libraries was measure using the Qubit HS DNA assay (Thermo Scientific) and the integrity was analysed with the Agilent 2100 Bioanalyzer using the DNA kit (Agilent, Waldbronn, Germany). All the libraries were pooled to be approximately equimolar.

#### 2.3.3.5 Illumina sequencing for RNA reads

The RNAseq libraries were pooled into one lane and run as paired-end 100 bp reads on the HiSeq platform using the SBS v4 chemistry. These aspects of the work were all conducted by the CGR (Liverpool).

#### 2.3.3.6 Assembly and mapping of Illumina reads

Prior to read mapping the reads, we used RSeQC (Wang, Wang, and Li 2012) to calculate the inner distance between two paired-end RNA reads. All *P. graminicola*'s reads were mapped to the genome using TopHat version v2.1.0 (Trapnell et al. 2009, 2012). For TopHat we change the default parameters for: library type, mate inner distance, maximum intron length and minimum intron length, which were set as *first strand*, *250 bp*, *5000 bp* and *20 bp*, respectively. Cufflinks version 2.2.1 (Trapnell *et al.* 2009) was used to assemble the transcripts with default parameters except for maximum intron length and library type, which were the same as those used for TopHat (Appendix. 2-). The assembled transcripts were used for annotation purposes.

### 2.3.4 *P. graminocola's* genome annotation

#### 2.3.4.1 Gene calling

Gene prediction utilising both RNAseq and genomic data was conducted with the *Braker* pipeline version 1.9 (Hoff *et al.* 2016) which implements two gene prediction tools, GeneMark-ET and Augustus. We used the list of coordinates of introns, provided from the Cufflinks output files (*junction.bed*). As instructed in the Braker 2.1.0 manual, we used the code *bed_to_gff.pl* from GeneMark–ET to transform the "*bed*" format from Cufflinks output to "*gff*" format. We then ran Braker with default parameters (Appendix. 2-).

#### 2.3.4.2 Functional annotation: Blast2Go Suite

Orthologous relationships between the *P. graminicola* proteins predicted by the Braker pipeline and other fungi species were evaluated by Blast2GO online ensuite (Götz et al. 2008) against the Swissprot non-redundant database (Bairoch et al. 2004). Blast2GO is a bioinformatic tool for automatic functional annotation based on Gene Ontology (GO) terms. It goes through three steps. The first one runs a typical NCBI-Blast, then maps the BLAST hits to the corresponding GO annotation from the chosen database and some additional data files within the tool (Figure 2-1).

#### 2.3.4.3 Functional annotation: InterProScan

To complement the GO analysis we used InterProScan 5 version 5.23-62.0(Jones et al. 2014) from the command line. Validity of the resulting annotation was assessed by looking to total length of match, the e-value and comparing the functional assignment to see how well they correspond. Finally, in order to filter all the hits obtained from InterProScan we ran two rounds of filtering. In the first round we selected as true hits those that have coverage greater or equal than 50% of the imputed sequence length, with a cut-off value of 0.05 or smaller. On the second round, we selected only those hits for which a function or structural domain was known, regardless of the presence or absence of a GO term or pathway annotation. From the previously selected hits we removed duplicated genes prior to checking their annotation (Figure 2-1).

#### 2.3.4.4 Functional annotation: KEGG & OrthoMCL

For those genes for which no annotation was accomplished by implementing the aforementioned pipeline, we complemented with KEGG and OrthoMCL (following the previously explained filtering) both from the online ensuite. The former using the online suite KOALA (KEGG Orthology And Links Annotation) specific to fungi

as taxonomy group; the latter was run from command line utilising the module 1 and the software version 4.1 against *U. maydis* proteome (Figure 2-1).

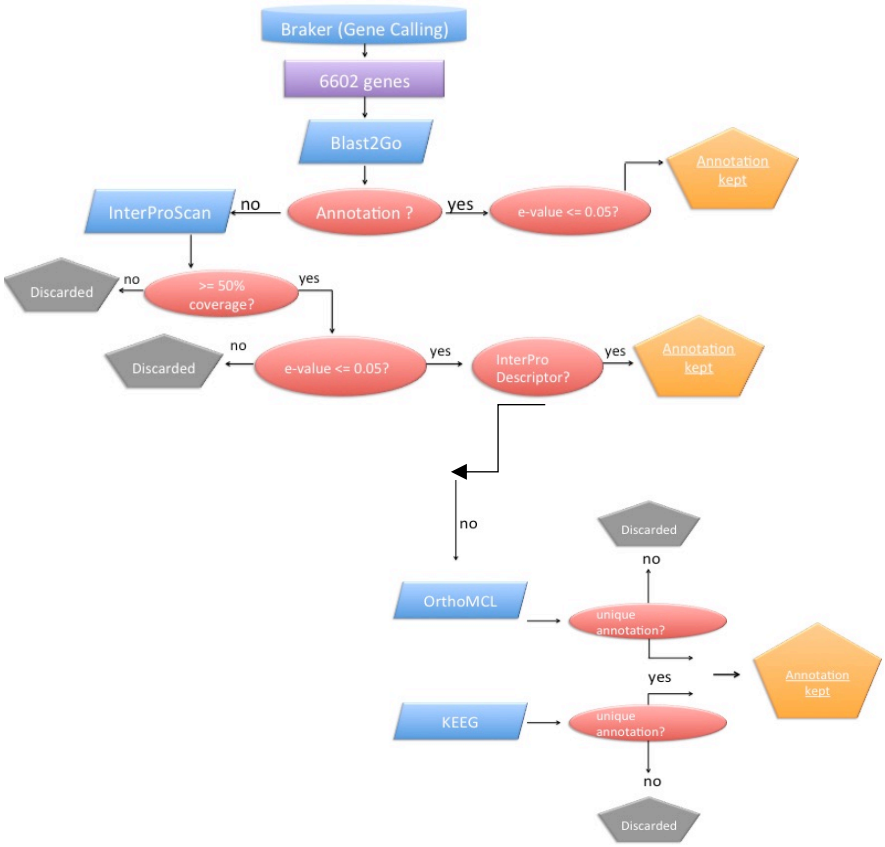



Figure 2-1. Diagram for filtering process implemented during functional annotation for *P. graminicola* genome.

#### 2.3.4.5 Evaluation of gene prediction

The gene prediction features from *P. graminicola* were calculated using the Eval package (Keibler and Brent 2003) and compared to *Ustilago hordei:* (accession number:GCA_000286035), *Ustilago maydis 521:* (accession number:GCF_000328475.2), *Pseudozyma hubeiensis SY62:* (accession number:GCA_000403515.1), *Sporisorium reilianum SRZ2:* (accession number:GCA_000230245.1), *Sporisorium scitamineum:* (accession

number:GCA_001010845.1), *Pseudozyma aphidis DSM 70725:* (accession number:GCA_000517465.1), *Aspergillus nidulans FGSC A4:* (accession number:GCA_000149205.2). For this we retrieved the *gtf* files from the EnsemblFungi website: http://fungi.ensembl.org/info/website/ftp/index.html. The *gtf* files were the input used by *eval.pl* script to get the corresponding comparisons.

#### *2.3.5* Comparative genomic analysis of *P. graminicola* and two pathogenic basidiomycetes

We used SignalP 4.0 from command line with default parameters to detect secreted proteins from *P. graminicola's* proteome. We assumed a protein was secreted if: it was predicted to have a secretion signal peptide, no transmembrane domain (based on SignalP) and if the protein started with methionine. These proteins were compared to effector clusters identified in plant pathogens such as *U. maydis* and other basidiomycetes.

## 2.4 RESULTS & DISCUSSION

### 2.4.1 PacBio sequencing: genome structure

In order to sequence the genome of *P. graminicola*, we utilized the PacBio technology as this potentially gives long reads of up to 10kb, which facilitates genome assembly. We prepared high quality genomic DNA with a fragment length in excess of 20 kb and this was utilised for sequence library production by the CGR (Liverpool) (Figure 2-2).

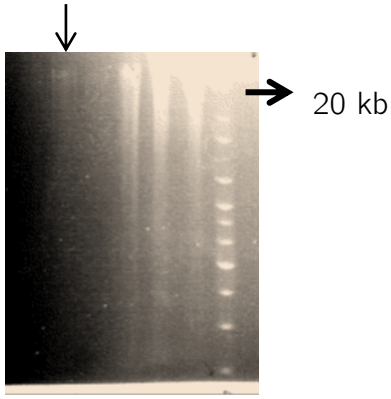



Figure 2-2. Integrity of gDNA extracted from *P. graminicola*. The DNA samples was evaluated using agarose gel electrophoresis. A DNA ladder was used as a marker. The genomic DNA (lanes 1-5) migrated above the 20 kb marker (indicated by a horizontal arrow).

The resulting sequence data (4 SMRT cells, 293,670 reads and average coverage of 157.52) was assembled by the Hierarchical Genome Assembly Process (HGAP) version 2.4.0 developed by PacBio to generate *de novo* assemblies.

The best assembly (complete and contiguous contigs) obtained resulted in 34 contigs with a total of 19.57 Mb of sequence. This is consistent with the expected genome size based on the genomes of related species such as *U. maydis* (19.7Mb), *S. reilianum* (18.5 Mb), *M. antarcticus* T-34, previously named *P. antarctica* T-34, (18.7 Mb) and *M. aphidis* DSM 70725, previously named *P. aphidis* DSM 70725 (17.9 Mb).

Our assembly into 34 contigs suggests we almost achieved chromosome level assemblies, when it is compared to species such as *U. maydis* (a very well studied organism), which has 23 chromosomes. The resulting assembly seems more complete and uniform than other those of related species currently available at the NCBI (Table 2-1), which present a large number of contigs.

Table 2-1. Assembly metrics for P. graminicola and related fungi

| Organism | Unit as reported on NCBI | Total length (Mb) | Total Genes |
|---|---|---|---|
| *P. graminicola* | 34 Contigs | 19.57 | 6602 |
| *S. reilianum* | 23 Chromosomes | 18.5 | 6806 |
| *U. maydis* | 23 Chromosomes | 19.7 | 6910 |
| *P. antarctica T-34* | 761 Contigs | 18.7 | 6560 |
| *P. aphidis DSM 7025* | 1 968 Contigs | 17.9 | 6011 |

### 2.4.2 RNA mapping reads

We then implemented transcriptome sequencing to improve the gene calling by identifying intron and exon positions and non-coding borders. For this we used the HiSeq platform and combined RNA extracted from samples grown under induced (presence of FA) and non-induced (without FA) conditions, supplemented across a 117 h time course.

Overall, we obtained a high number of reads mapped, which range from 96.40% to 98.20%, from which a low percentage showed multiple alignments (Table 2-2). Additionally, we detected a total of 7190 transcripts.

Table 2-2. Mapped reads summary for *P. graminicola* fermenter sample

| Time-point (h) | Condition | Mapped reads (bp) | Mapped reads (%) | Multiple alignment reads (%) | Multiple alignment reads (%) | Total (bp) |
|---|---|---|---|---|---|---|
| 24 | non-induced | 34,145,633 | 97.8 | 2 | 5.4 | 34,923,608 |
| 24 | non-induced | 315,375 | 97.2 | 3 | 11.1 | 324,534 |
| 48 | non-induced | 50,417,756 | 97.9 | 2 | 3.1 | 51,512,093 |
| 48 | non-induced | 17,640,956 | 98.1 | 2 | 5.2 | 17,973,607 |
| 72 | non-induced | 20,022,098 | 97.9 | 2 | 4.8 | 20,459,560 |
| 72 | non-induced | 3,936,638 | 97.9 | 2 | 11.4 | 4,022,760 |
| 96 | non-induced | 1,584,033 | 97.3 | 3 | 10.4 | 1,627,288 |
| 117 | non-induced | 277,539,445 | 97.3 | 3 | 8 | 285,364,276 |
| 24 | induced | 147,101,732 | 98.2 | 2 | 2.7 | 149,831,881 |
| 24 | induced | 3,962,473 | 97.9 | 2 | 4.6 | 4,046,659 |
| 48 | induced | 59,700,123 | 98 | 2 | 3.7 | 60,892,928 |
| 48 | induced | 8,492,731 | 97 | 3 | 6.2 | 8,759,067 |
| 72 | induced | 25,721,066 | 97.9 | 2 | 5.7 | 26,279,038 |
| 72 | induced | 9,249,548 | 97.2 | 3 | 7 | 9,512,680 |
| 96 | induced | 1,954,727 | 96.4 | 4 | 5.9 | 2,028,150 |
| 117 | induced | 23,944,220 | 97.6 | 2 | 5.9 | 24,520,473 |

### 2.4.3 Overall *P. graminicola* genomic features

Among the genomic features used to determine the quality of an assembly, the N50 refers to the length where 50% of all bases are contained in sequences longer than this value. Therefore, the higher the N50 the better the contiguity and completeness of a genome assembly. Our N50 is larger than the average sequence length (Table 2-3), which suggests we got complete genes on the assembled contigs. This is supported as well by the lack of gaps (Ns) through the whole assembly (Table 2-3).

The %GC content for *P. graminicola* is similar to those of closely related species, such as *U. maydis 521* (53.9 %), *S. reilianum SRZ2* (59.5 %) and *Pseudozyma hubiensis SY62* (56.6 %).

Table 2-3. Genomic features for *P. graminicola* assembly

| Features | *P. graminicola* |
|---|---|
| %GC Content | 56.75 |
| Longest Contig Length (Mb) | 2.35 |
| Min sequence length | 1702 |
| Max sequence length | 2,352,515 |
| Average sequence length | 575,504.41 |
| Median sequence length | 555,394 |
| N50 Contig Length (bp) | 823,447 |
| Ns | 0% |

### 2.4.4 Phylogenetic analysis

In order to improve the genome assembly from our 34 contigs we ran a preliminary phylogenetic analysis, using the nucleotide sequences of *U. maydis*, *P. antarctica* and *P. aphidis*. We used the assembled genome and implemented progressive Mauve to align the genomes using default parameters. The goal of this was to identify the best reference (therefore closely related species to *P. graminicola*). On a first analysis we identified *U. maydis* more closely related to *P. graminicola*

than other *Pseudozyma* species, for which we repeated the analysis and included more species. We observed that *S. reilianum* is more closely related to *P. graminicola* than the other two *Pseudozyma* species (Figure 2-3), consequently both *U. maydis* and *S. reilianum* could work as good references to improve the arrangement of the genomic assembly.

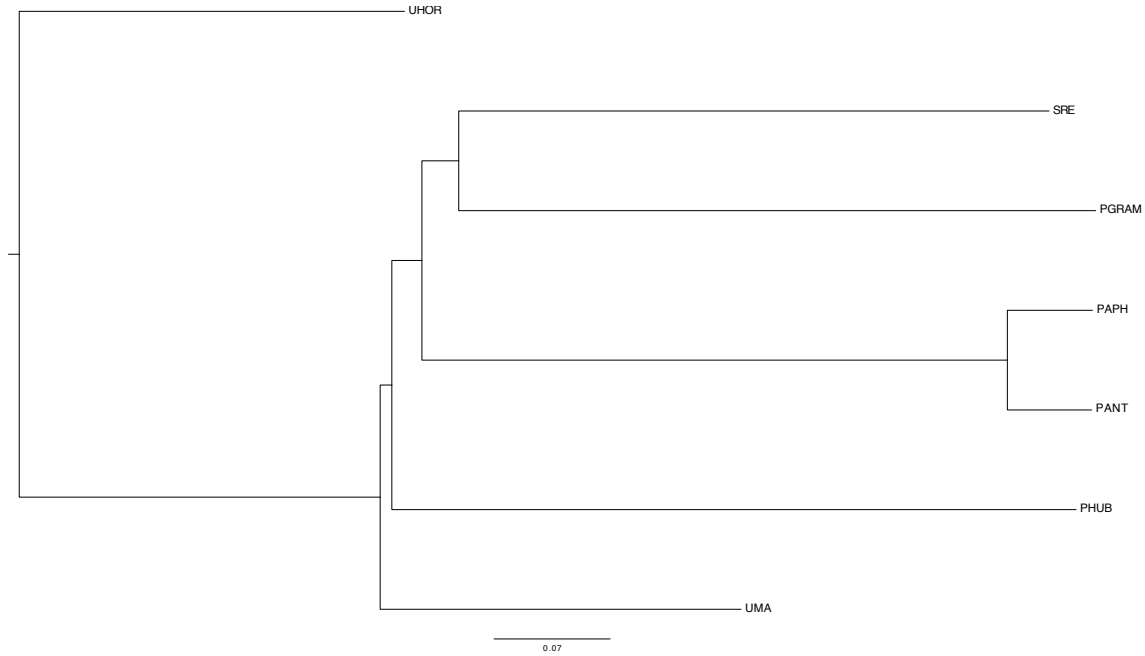



Figure 2-3. Molecular phylogenetic tree constructed using the nucleotide sequence for *P. graminicola* and other related fungi. Species key: UMA: *Ustilago maydis* 521, PGRAM: *Pseudozyma graminicola*, SRE: *Sporisorium reilianum* SRZ2, PANT: *Moesziomyces antarcticus* T34, PAPH: *Moesziomyces aphidis* DSM 70725. The alignment was done using progressive mauve (default parameters). The tree was drawn using FigTree.

### 2.4.5 Arrangement of *P. graminicola's* assembled genome

From the preliminary phylogenetic analysis, we identified *U. maydis* and *S. reilianum* as the optimal comparators for the *P. graminicola's* genome. For this we used MAUVE to align and rearrange the contigs. The comparison to *U. maydis* revealed a high level of synteny with *P. graminicola* genome (Figure 2-4), with extended regions of unbroken homology across entire *P. graminicola* contigs. Additionally, some inversions (blue lines) and translocations are evident, one

translocation comprising a section equivalent to *U. maydis* chromosome 3 and the entirety of chromosome 8. In general, the pairwise comparison between both genomes tends to follow a straight diagonal line, which suggests both the genome content and organization are very similar between the two species.

The equivalent comparison of *P. graminicola* to *S. reilianum* also confirmed a close relationship (Figure 2-5)

Again, the alignment shows a few inversions in *P. graminicola's* genome, with respect to *S. reilianum* (blue lines), particularly in chromosomes 5, 8, 12 and 17. As well as with the *U. maydis* comparison, homology between the two genomes tends towards a straight diagonal line, which supports the phylogenetic clustering of *P. graminicola* with these two species.

A comparison between the two reference genomes was also conducted (Figure 2-6). As a result of this comparison an almost straight line is evident with only one major re-arrangement with a section of *U. maydis,* chromosome 5 aligning with chromosome 20 of *S. reilianum*. This analysis confirms that both these reference genomes share a very similar genetic arrangement consistent with either representing a good training set for *P. graminicola* annotation and analysis (Schirawski et al. 2008). However, *U. maydis* is a model plant pathogen which has been extensively studied (Hewald et al. 2006; Nugent, Choffe, and Saville 2004; Teichmann et al. 2007), and consequently it has a better genome annotation than *S. reilianum*. For this reason, we chose *U. maydis* as the primary reference genome.

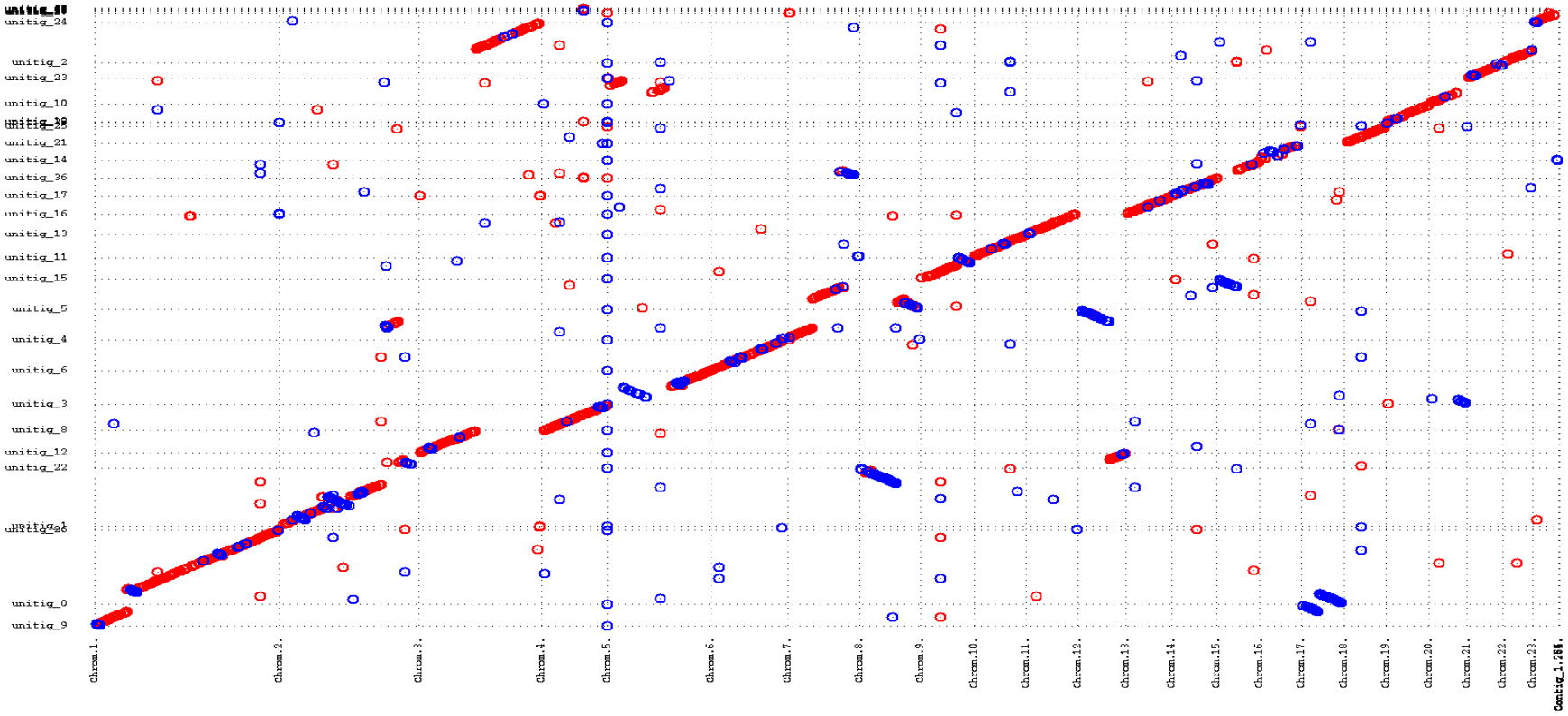



Figure 2-4. Comparison of the *P. graminicola* and *U. maydis* genomes. Visual display comparing for the genomic structure produced by the NUCmer tool and MUMmerplot. Regions of similarities between genomes are shown as straight diagonal lines. Blue indicates inversion of the corresponding fragment (according to the reference). The X-axis corresponds to reference genome, *U. maydis* and the Y-axis corresponds to the *P. graminicola* genome assembly.

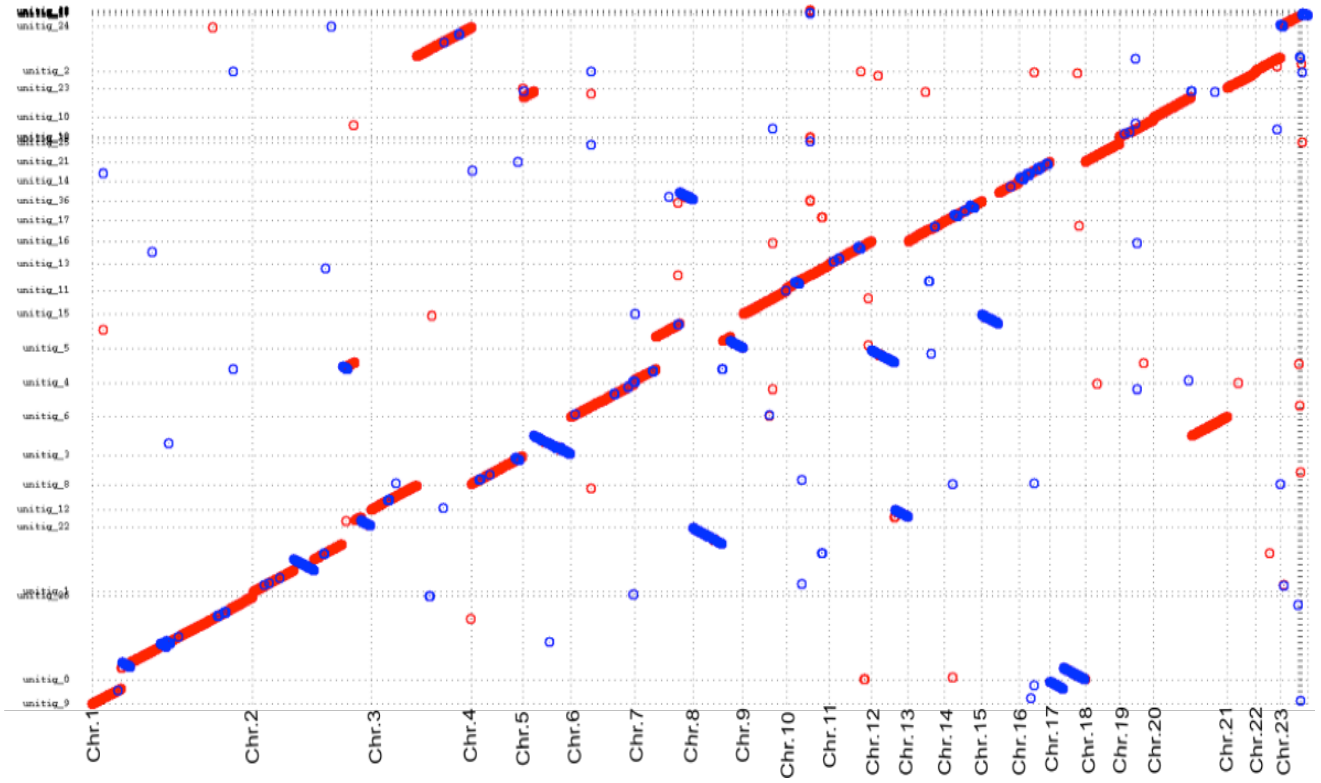



Figure 2-5. Comparison of *P. graminicola* and *S. reilianum* genomes. Visual display comparing for the genomic structure produced by the NUCmer tool and MUMmerplot. Regions of similarities between genomes are shown as straight diagonal lines. Blue indicates inversion of the corresponding fragment (according to the reference). The X-axis corresponds to reference genome, *S. reilianum* and the Y-axis corresponds to the *P. graminicola* genome assembly.

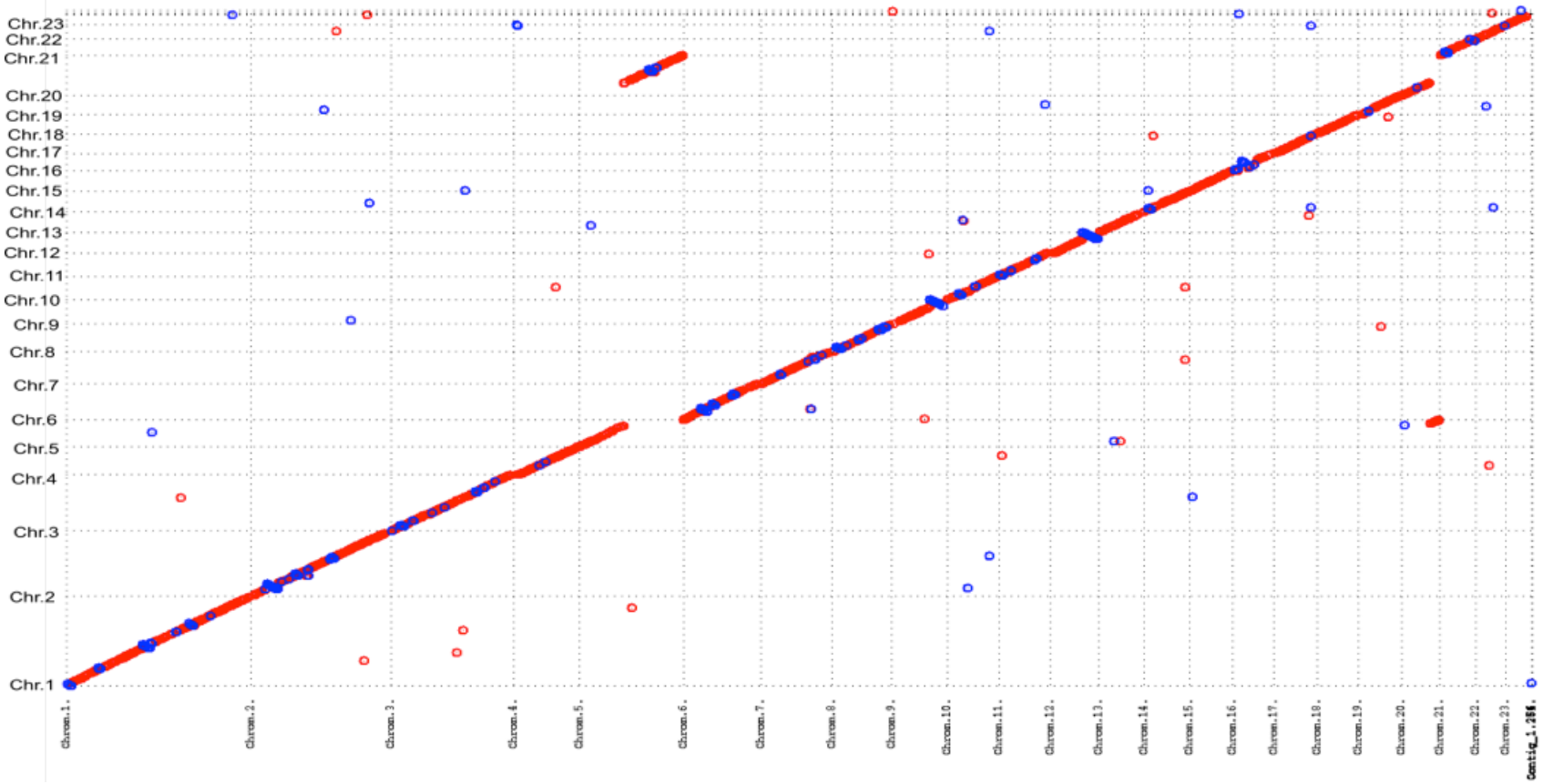



Figure 2-6. Comparison of *U. maydis* and *S. reilianum* genomes. Visual display comparing for the genomic structure produced by the NUCmer tool and MUMmerplot. Regions of similarities between genomes are shown as straight diagonal lines. Blue indicates inversion of the corresponding fragment (according to the reference). The X-axis corresponds to reference genome *U. maydis* and Y-axis corresponds to the *S. reilianum* genome.

### 2.4.6 Gene Calling

A key motivation in sequencing the genome of *P. graminicola* was to identify key genes associated with MEL biosynthesis, including the biosynthetic cluster. Also to identify genes involved in the pathway that will be required for or may influence MEL production, including regulatory genes such as those encoding transcription factors. For this reason, a full analysis of the coding potential of *P. graminicola* was required. In order to do this initially we combined RNA-seq and genomic sequence data together with the Braker pipeline. Braker implements GeneMark-ET (Lomsadze et al. 2014) which incorporates RNA-Seq reads into the prediction training. From this prediction, a subset of genes having support from the RNA-seq alignments for their structure, including all intron boundaries are used to train Augustus. At this stage, Augustus will perform the gene calling using the RNA-seq evidence provided; therefore, the prediction will be more accurate than using only the genomic data.

A total of 6602 predicted genes were defined using the Braker pipeline. As this is the first gene prediction for *P. graminicola* we compared our results to a set of closely related fungi (this selection was based on availability of genome data of MELs producers). The most common measurements to determine the performance of a gene prediction are sensitivity and specificity. Sensitivity relates to the coverage of coding features (nucleotide/exons) called, compared to the reference. Specificity defines the accuracy of those features, as an example, for a coding region to be called properly, both boundaries (5' and 3' ends) have to be predicted correctly, when compare to a reference (Guigó et al. 2000). These measurements provide an insight of how well the annotation system has performed. These values must be interpreted in the relative context as they might indicate different things according to the sequence data and the reference used.

For example in a case were sensitivity is being measured at the transcript level, if a poor performance is observed for long genes but a good performance for short genes, this may relate to divergence in intron length or exon count (Keibler and Brent 2003). The specificity on the other hand is a reflection of the accuracy of the called feature.

Under these definitions, overall, we found higher values for sensitivity over specificity, which indicates a high proportion of features (coding nucleotides, exons, transcripts) being correctly predicted in our gene set. The concordance between specificity and sensitivity is a measure of relationship between species, as closely related species are likely to have similar features. Additionally, the variation in this category, at nucleotide level, is likely to occur due to splicing, indels or mutations that might shift the ORF of genes (Table 2-4).

In our prediction, both overall scores, for sensitivity and specificity at the three levels of the gene prediction showed the highest values when compared to *U. maydis* (Table 2-4). This result is also supported by the high level of synteny shared between both strains; hence the gene architecture is likely to be similar to that of *U. maydis*'s.

At the transcript level, as expected, most of the metrics are similar between all the species, excluding *A. nidulans,* which was used as an out group and therefore expected to have marked differences for values such as transcript count, exons per gene and total of exons (Table 2-4). Interestingly the exon length in all the basidiomycetes is high, compared to *A. nidulans,* an ascomycota.

The concordance for specificity and sensitivity values for intron category shows the likeness between *P. graminicola,* S. *reilianum* and *P. hubeiensis*.

Table 2-4. Evaluation of Gene prediction from Braker pipeline compared to other *Basidiomycetes*

| LEVEL | FEATURE | *PGRA* | *SCI* | *SRE* | *PHUB* | *UMA* | *UHOR* | *PAPH_DSM* | *ANID* |
|---|---|---|---|---|---|---|---|---|---|
| Gene | Count | 6602 | 7711 | 6803 | 7619 | 6910 | 7230 | 6011 | 9977 |
| | Total Transcripts | 7191 | 7712 | 6804 | 7620 | 6929 | 7232 | 6012 | 9978 |
| | Transcripts p/gene | 1.09 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | Genomic Overlap Specificity | - | 99.57% | 79.11% | 86.40% | 94.79% | 49.72% | 79.91% | 78.14% |
| | Genomic Overlap Sensitivity | - | 1.72% | 66.85% | 99.80% | 96.54% | 98.33% | 66.71% | 100% |
| | CDS Overlap Specificity | - | 4.79% | 94.12% | 86.25% | 94.78% | 49.72% | 93.58% | 77.95% |
| | CDS Overlap Sensitivity | - | 97.81% | 95.27% | 97.89% | 94.66% | 98.33% | 97.04% | 95.63% |
| Transcript | Count | 7191 | 7712 | 6804 | 7620 | 6929 | 7232 | 6012 | 9978 |
| | Average Length | 1895.11 | 1568.32 | 1820.43 | 1833.22 | 1799.26 | 1754.15 | 1915.51 | 1791 |
| | Median Length | 1587 | 1217 | 1504 | 1482 | 1479 | 1420 | 1581 | 1502 |
| | Total Length | 13627741 | 12094908 | 12386195 | 13969170 | 12467083 | 12686021 | 11516056 | 17879759 |
| | Average Coding Length | 1828.76 | 1458.12 | 1752.79 | 1633.07 | 1731.71 | 1673.83 | 1783.47 | 1523 |
| | Median Coding Length | 1518 | 1119 | 1440 | 1305 | 1428 | 1332 | 1488 | 1275 |
| | Total Coding Length | 13150593 | 11245031 | 11926008 | 12443985 | 11999008 | 12105117 | 10722201 | 15203067 |
| | Average exons p/gene | 1.46 | 1.7 | 1.43 | 1.44 | 1.41 | 1.52 | 1.84 | 3.52 |
| | Total Exons | 10521 | 13090 | 10521 | 10953 | 9744 | 10992 | 11041 | 35108 |
| Exon | Count | 10444 | 13073 | 9750 | 10953 | 9732 | 10990 | 11041 | 35106 |
| | Average Length | 1253.77 | 859.73 | 1223.13 | 1136.13 | 1231.91 | 1101.42 | 971.13 | 433 |
| | Median Length | 903 | 426 | 858 | 675 | 864 | 609 | 510 | 217 |
| | Total Length | 13094374 | 11239248 | 11925502 | 12443985 | 11988940 | 12104609 | 10722201 | 15202686 |
| | Overlap Specificity | - | 4.85% | 92.73% | 84.70% | 93.43% | 48.14% | 91.56% | 76.02% |
| | Overlap Sensitivity | - | 95.33% | 95.99% | 99.65% | 95.84% | 99.95% | 95.50% | 99.88% |
| Intron | Count | 3292 | 5375 | 3078 | 3481 | 2953 | 3880 | 5030 | 25555 |
| | Average Length | 143.36 | 154.1 | 135.81 | 3481 | 145.58 | 140.91 | 154.24 | 101.57 |
| | Median Length | 95 | 99 | 87 | 203 | 100 | 98 | 95 | 63 |
| | Total Length | 471945 | 828265 | 418031 | 1482680 | 429890 | 546736 | 775821 | 2595739 |
| Nucleotide | Overlap Specificity | - | 4.07% | 27.37% | 48.78% | 27.61% | 29.07% | 41.40% | 56.90% |
| | Overlap Sensitivity | - | 37.38% | 28.78% | 48.41% | 29.33% | 36.68% | 30.68% | 33.96% |
| | Specificity | 67.61% | 1.72% | 66.85% | 47.35% | 79.03% | 99.96% | 66.71% | 98.22% |
| | Sensitivity | 79.06% | 99.57% | 79.11% | 95.15% | 67.63% | 18.14% | 79.91% | 36.47% |

Species key: PGRA: *P. graminicola*, SCI: *Sporisorium scitamineum*, SRE: *Sporisorium reilianum* SRZ2, PHUB: *Pseudozyma hubiensis* SY62, UMA: *Ustilago maydis* 521, UHOR: *Ustilago hordei*, PAPH_DSM: *Pseudozyma aphidis* DSM 70725, ANID: *Aspergillus nidulans.*

The discrepancy for *P. graminicola* gene count and transcript count values could be attributed to the fact that multiple transcripts could belong to the same gene, therefore the count for transcripts is expected to be higher than the gene count as a gene might include more than one transcript (Figure 2-7.).

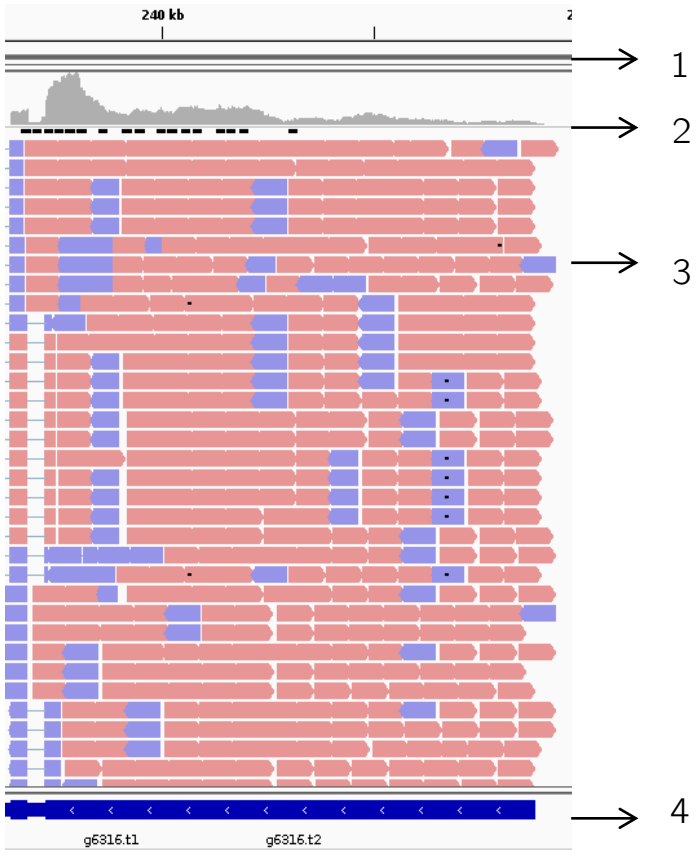



Figure 2-7. IGV visualisation for a zoomed-in section of *P. graminicola* genome. 1) Top panel corresponds to contig displaying localisation in kilo bases. 2) RNA-seq coverage for the zoomed-in gene. 3) Aligned reads corresponding to forward (pink colour) or reverse (blue colour) orientation. 4) Multi-exons for the gene 6316, displayed as transcripts.

### 2.4.7 Functional Annotation: integration of tools

In order to get the functional annotation for *P. graminicola* genome, we used the full complement of genes identified using the Braker pipeline with the Blast2Go suite and the Swissprot database, restricted to fungi species. For those genes

without an annotation we extended the search to InterProScan, implementing filters regarding length of coverage and e-value. Next, for the genes without an InterProScan descriptor we implemented OrthoMCL and KEGG tools, in order to get an annotation for those genes (Figure 2-1).

From the first analysis, Blast2Go assigned a function or description to more than 65% of the putative gene sequences, OrthoMCL clustered more than 82% of the sequences and KEGG only provided terms for approximately 27% of the sequences (Table 2-5). For OrthoMCL 20 genes did not clustered to any group, when compared to *U. maydis* proteome.

Table 2-5. Overall statistics for the gene annotation of P. graminicola genome.

| | Blast2Go | KEGG |
|---|---|---|
| Annotated | 4492 | 1750 |
| Non-annotated | 2110 | 4852 |
| Total | 6602 | 6602 |

Values based on unique gene entries and/or unique annotation/description entries

In order to assess the extended annotation obtained with InterProScan databases we analysed the data provided by each database individually and looked for those with a coverage higher than 50% as these hits will be expected to be more reliable than those covering a smaller percentage of the protein (Korf et al. 2003) (Table 2-6).

Table 2-6. Statistics for InterProScan results using the *P. graminicola* gene level annotation

| Database | 3/3 | 1/3 | Unique |
|---|---|---|---|
| CDD | 114 | 353 | 12 |
| HAMAP | 197 | 260 | 7 |
| PANTHER | 639 | 1343 | 7 |
| PFAM | 44 | 1943 | 7 |
| PRODOM | 17 | 24 | 1 |
| PROSITEPATTERNS | 1 | 1 | 0 |
| PROSITEPROFILES | 74 | 530 | 44 |
| SMART | 78 | 318 | 29 |
| SUPERFAMILY | 79 | 1406 | 184 |
| TIGRFAM | 290 | 448 | 10 |

**3/3**: the annotation fulfils three of the three requirements, which were: to have an InterPro description, a GO term and a pathway annotation. **1/3**: the annotation has at least an InterPro description but not necessarily a GO term nor pathway annotation. **Unique** = the number of hits identified with that database that were not found with Blast2Go and are therefore consider unique for that particular database.

The databases, which provided the most unique hits, were SUPERFAMILY, ProSiteProfiles and SMART, respectively. SUPERFAMILY database contains structural and functional annotation for proteins from completely sequenced genomes, including 173 fungal species from which the *Basidiomycete* group is represented by *U. maydis* and *S. reilianum*, both of which are closely related to *P. graminicola*. This database groups together domains from known structures, which have an evolutionary relationship (Murzin et al. 1995). On the other hand, ProSiteProfiles compares the amino acid sequences to a matrix of multiple sequence alignments, which are used to create a frequency distribution of amino acid occurrence (Sigrist et al, 2002). ProSiteProfiles annotation tries to cover the entire length of a protein family or domain instead of only a small region with high sequence similarity. This could lead to false positives; in order to verify a hit, it has to "correctly align those residues having analogous functions or structural properties according to experimental data" (Sigrist *et al*. 2012) which means the

annotation can be extended to proteins with a weaker evolutionary relationship. SMART database also performs sequence alignments and seeks genetically mobile domains, utilising domain architectures, that have been manually curated, which in our case led to a low number (Ponting *et al*. 1995).

This procedure gave us an annotation distributed as follows: 4492 unique genes annotated/described with Blast2Go (although some annotations included the word: "uncharacterised", "putative" or "possible"), 301 unique genes annotated with InterProScan, 51 with KEGG and 152 with OrthoMCL. This gives a total of 4996 unique genes annotated for *P. graminicola*, which corresponds to approximately a 76% of the genome annotated on the first attempt.

On a second attempt, we analysed the data from a more integrative approach. For this, we clustered the InterPro results by its InterPro annotation (e.g. IPR002347) and type of classification (family, domain, homologous superfamily, repeat, and sites) regardless the database used, and compared the number of annotated genes to the Blast2Go output (for the purpose of this analysis KEGG and OrthoMCL were excluded as individual database, to avoid redundancy). As a result, InterProScan allowed us to annotate a total of 5487 genes whereas Blast2Go annotated 4492. With such integration our final annotation increased from 4996 genes (first analysis) to 5547, representing an 84% of the total genome annotated. We kept this annotation and calculated the number of entry types obtained for each gene and used SignalP to predict secreted proteins (Table 2-7)

Table 2-7. Type of entry obtained with InterPro Scan analysis and SignalP for P. graminicola genome annotation.

| Type of Entry | Proteins* |
|---|---|
| Active Site | 218 |
| Binding Site | 170 |
| Conserved site | 786 |
| Domain | 3825 |
| Family | 2743 |
| Homologous superfamily | 2326 |
| PTM | 27 |
| Repeat | 306 |
| Secreted proteins | 508 |

*These values can be redundant; a gene can match more than one type of entry

The number of predicted secreted proteins expressed by *P. graminicola* (508), is relatively high compared to those reported for species such as *U. maydis* (467) and *S. reilianum* (467) (Schuster et al 2016). This might be due to the pipelines we used for gene calling and annotation, as our selection of parameters might differ from the ones used during the annotation of the species we are comparing to.

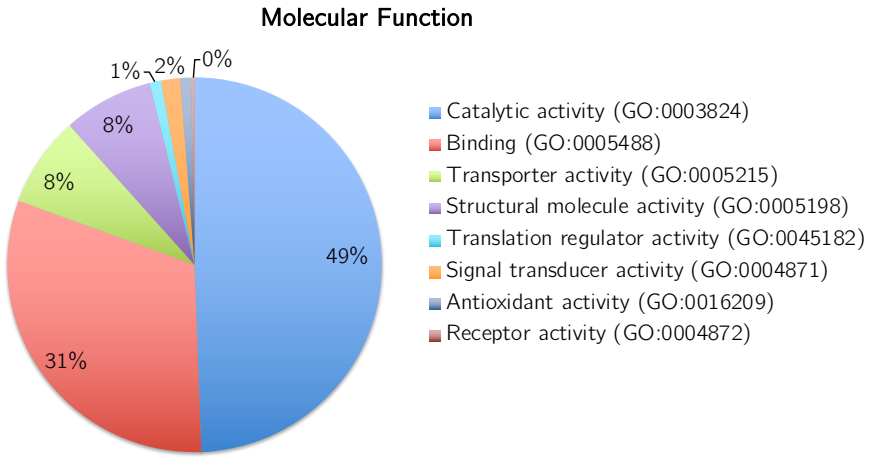
### 2.4.8 Comparative genomic analysis

#### 2.4.8.1 P. graminicola Gene Ontology (GO) based distribution of genes

Despite the lack of biological information for *P. graminicola's* ecology, we aimed to attribute some insight to it by comparing the proteome of this smut fungus to that of *U. maydis*. From such comparison we investigated the distribution of the functional categories, based on Gene Ontology (GO) from PANTHER-GOslim, derived from homologous proteins. We selected only hits which had a significant p-value ($p < 0.05$) and a percentage of identity equal or higher than 50%. By implementing these parameters, we got 4013 hits to *U. maydis* proteome, corresponding to 65% of transcripts mapped to *U. maydis*

(http://www.pantherdb. org/, last accessed March 13, 2018). Allocation by GO domains classified as "Molecular Function" (MF) and "Biological Process" (BP) are represented in Figure 2-8, A and B, respectively.

A)




B)

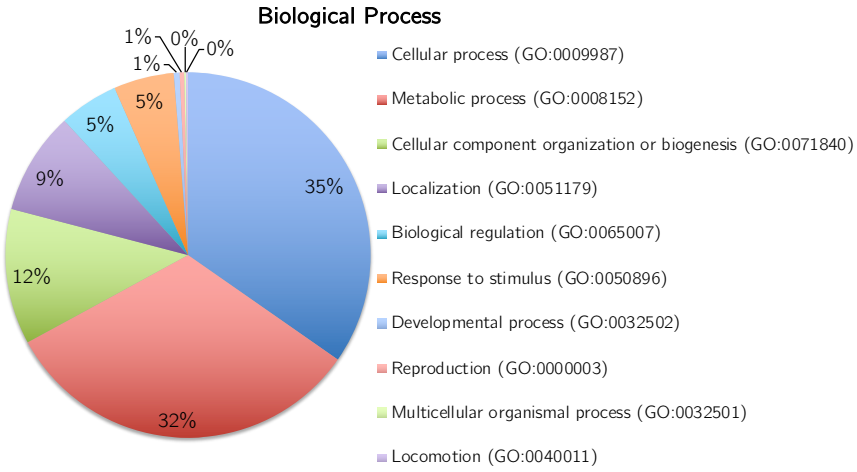



Figure 2-8. GO-slim categorical designation for functional annotation of genes present in the *P. graminicola* proteome obtained from PANTHER online ensuite by homology to *U. maydis* proteome. A) Molecular function and B) Biological process domains presented.

We also looked for the main protein classes encoded by *P. graminicola* genome, and found nucleic acid binding, hydrolases, transferases and oxidoreductase, being the most abundant (Figure 2-9). Among the hydrolase class, *P. graminicola* has

15 genes encoding for enzymes of the type glycoside hydrolase family 16 (Appendix. 2-). This family has a carbohydrate-active enzyme (CAZymes) classification (Cantarel et al. 2009), which has been associated, in fungi, to encode host-targeted, hydrolytic enzymes acting on plant bio-polymers (Cantarel et al. 2009). Consistent with these findings, among the transferase enzyme group we also identified enzymes with CAZyme classification as enzymes related to oxidation-reduction interaction (Appendix. 2-). The later interaction is further represented by the oxidoreductase group, which is also abundant in *P. graminicola* proteome (Figure 2-9). An important protein family involved in processes such as defence and virulence mechanisms, browning and pigmentation and melanin production are the glucose-methanol-choline (GMC) oxidoreductases due to its catalytic activity which transfer electrons between molecules (Xu et al. 2016). We found genes coding for enzymes associated to this enzyme classification (EC) (GMC oxidoreductases) and identified some enzymes associated to choline substrate and two genes encoding members of the FAD/NAP(P)-binding domain superfamily.

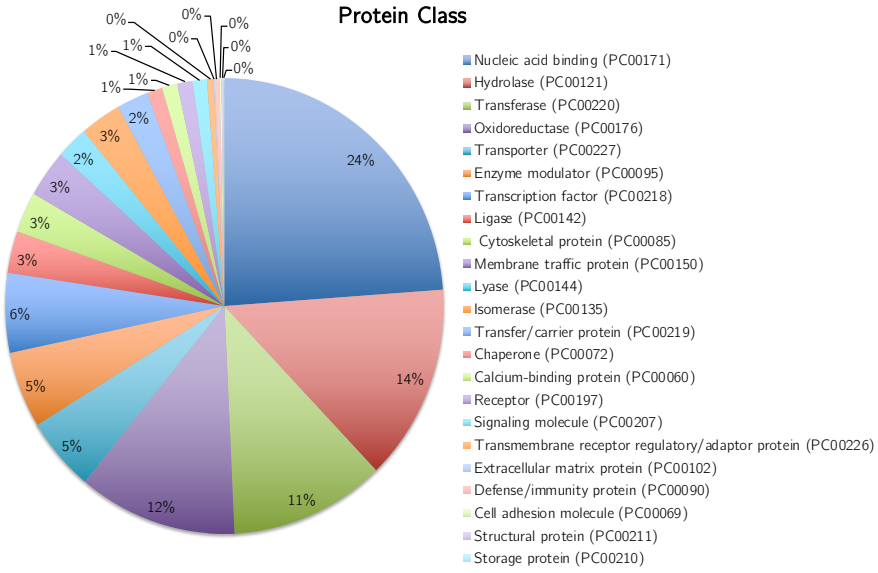



Figure 2-9. Principal classes of protein encoded by *P. graminicola* genome.

As an addition to these findings we looked specifically for pathogenicity associated genes reported in other smut fungi. We found *P. graminicola* carrying most of them; and in similar numbers as other related pathogenic species (Sharma *et al*. 2015) (Table 2-8).

Table 2-8. Candidate for putative pathogenicity enzymes in four smut fungi.

| Function | *U. maydis* | *S. reilianum* | *M. pennsylvanicum* | *P. graminicola* |
|---|---|---|---|---|
| ATP-binding cassette/(ABC) transporter | 22 | 19 | 19 | 19 |
| Protease inhibitor | 3 | 4 | 4 | 2 |
| Phospholipase | 11 | 12 | 12 | 15 |
| Lipase | 33 | 33 | 32 | 36 |
| Cysteine protease inhibitor | 1 | 1 | 1 | 1 |
| Serine protease | 51 | 54 | 46 | 46 |
| Aspartic protease | 11 | 10 | 10 | 8 |
| Glycosyl hydrolases | 27 | 31 | 23 | 18 |
| Cutinase | 4 | 3 | 2 | 4 |
| Pectin esterase | 1 | 1 | 1 | 1 |
| Cytochrome P450s | 21 | 16 | 13 | 17 |
| Pectin lyase fold/virulence factor | 4 | 5 | 4 | 1 |

### 2.4.9 Secreted proteins in *P. graminicola*

We have produced the first genome sequence and annotation for *P. graminicola*, a smut fungus closely related to *U. maydis*. *P. graminicola* was isolated from herbaceous plants (Golubev et al 2007) and its infective process has not been described yet, but our findings so far suggested a high relatioship to pathogen species, hence we seeked for orthologs to effectors described for *U. maydis.* Effectors are secreted pathogen proteins which suppress host defences or tune host metabolism to support the infection process (Lanver et al. 2017).

In *U. maydis* the infection process and therefore effectos have been well described and characterised. However, this smut fungi comprises 467 secreted proteins, from which 203 are completely novel. Additionally, smut fungi related to *U. maydis* have similar number of predicted secreted proteins (Lanver et al. 2017).

To investigate this further, we focused on the main effectors described for *U. maydis*, which have been extensively studied, the 19A and other secreted proteins (Brefort et al. 2014; Kämper et al. 2006; Schuster et al 2016; Tanaka et al. 2014).

In *U. maydis* after the dikaryon is formed (fusion of hyphae) penetration structures are developed aided by secretion of plant cell wall degrading enzymes; which are not many in smut fungi (Lanver et al. 2017) highlighting their biotrophic nature. Nevertheless, the biotrophic interaction inside the host is mediated by effector proteins which facilitate the nutrition and regulate the process and events of the infection process. Among the main effectors described in *U. maydis* Pep1 inhibits the plant peroxidase (Hemetsberger et al., 2012), Pit2 inhibits apoplastic cysteine proteases (Mueller et al., 2013). Cmu1 from *U. maydis* alters the chorismate homeostasis (Djamei et al., 2011). Tin2 has an effect which ends up in affecting lignin biosynthesis, altering fungal spread in the host tissue (Tanaka et al., 2014). See1 is required for the reactivation of plant DNA synthesis after infection. *U. maydis* and members from the smut family are pathogens to important worldwide crops (Brefort et al. 2014; Doehlemann et al. 2011; Kämper et al. 2006; Tanaka et al. 2014; Taniguti et al. 2015; Tollot et al. 2016; Wollenberg and Schirawski 2014), therefore increasing the current knowledge of mechanisms acting on the trigger and regulation of effectors in the vast majority of smut fungi is imperative.

We took the genes coding for the main effectors from *U. maydis* and blasted them against the *P. graminicola* genome. Among these, the largest effector cluster described in *U. maydis*, the cluster 19A comprises 24 genes. This is plant-induced and the expression of some of the genes is tissue-specific (Brefort et al., 2014). We decided to carry out this comparison as interestingly, we only identified eight

orthologs for the 19A cluster genes, at a cut-off of e-value $1e^{-5}$ (Table 2-9). Brefort and co-workers (2014) divided the 40kb 19Acluster into four regions and identified the genes that contribute the most to the virulence phenotype in *U. maydis* by deletion analysis. Each of the mutants generated was assayed for tumor formation and the ability to accumulate anthocyanin, as these two phenotypes are characteristics of the infection process (Kämper et al., 2006, Lanver et al., 2017). *P. graminicola* has orthologs mainly for the region 19A-1b and 19A-2 (Table 2-9). The section 19A-2 is a weak contributor to tumor formation in *U .maydis* and *S. reilianum*. Deletion of *S. reilianum tin4* and *tin5,* orthologs of the *U. maydis* effectors UMAG_05318 and UMAG_05319 respectively, weakly affected its virulence. We found an ortholog for *tin5* in *P. graminicola* (Table 2-9) and a two potential paralogs for *tin4* (Brefort et al., 2014).

The 19A-1b section from the cluster seems to contribute overall to the suppression of basal host immunity and encodes for the *tin1-1* to *tin1-5* effectors in *U. maydis* (Brefort et al. 2014). For this region we found orthologs for three out of the five genes comprising the cluster. In the case of *tin1-5* we found two paralogs (Table 2-9). These are proposed to function as avirulence proteins, which are under high selective pressure (Schuster et al 2016).

Although, *P. graminicola* has not been reported as a pathogen, it has an orthologue of the effector UMAG_10556 (*tin3*), g256, which shares 31% amino acid identity (Table 2-9). *tin3* is the major contributor to virulence in *U. maydis* and may be involved in tumour formation (Brefort et al. 2014). Additionally, this gene is not present in closely related pathogens that do not form tumour such as *S. reilianum* (Brefort et al. 2014). We did not find an ortholog in *P. graminicola* for the effector *tin2* (UMAG_05302), which is the major contributor to anthocyanin accumulation in *U. maydis* (Brefort *et al.* 2014) and is implicated in the reduction of lignin production in plants as a defence mechanism.

On the other hand, we found two paralogs for *tin4* (Table 2-9), which might indicate redundancy in the function of these genes. Although, this is only a theory as a deletion mutant would be required to confirm for this.

A reason for which we found such a low number of orthologs for the 19A cluster in *P. graminicola* is its relation to grass, as it was isolated from timothy grass -*Phleum pratense L.*- and meadow fescue -*Festuca pratensis Huds.* - (Golubev, et al 2007). Therefore, the mechanisms underling the gene expression of the plant during infection are different as has been shown for this cluster its expression is tissue specific and directly related to its different stages (Wollenberg and Schirawski 2014), supporting a specialisation on gene-host interaction. Hence, the gene function reported for *U. maydis* on maize (Kämper et al., 2006) not necesarilly need to be followed in *P. graminicola.*

Another factor affecting the orthologs found for the 19A cluster might be the cut-off criteria we used for the blast hits, not allowing us to detect homologous relationships.

Table 2-9. *P. graminicola* orthologs for effectors identified in *U. maydis*

| *U. maydis* ID | *P. graminicola* ID | % of identity |
|---|---|---|
| UMAG_05294[b] | ND | NA |
| UMAG_10554[b] | g253,g251 | 45.9, 30 |
| UMAG_05295[b] | g253 | 28.32 |
| UMAG_12302[b] | ND | - |
| UMAG_10553[b] | g252 | 36 |
| UMAG_05299[a] | ND | - |
| UMAG_05300[a] | ND | - |
| UMAG_05301[a] | ND | - |
| UMAG_05305[c] | ND | - |
| UMAG_05306[c] | ND | - |
| UMAG_10556[c] | g256 | - |
| *UMAG_05309 | ND | 45.88 |
| *UMAG_05310 | g258 | - |
| *UMAG_05311 | ND | - |
| *UMAG_05312 | g259 | - |
| *UMAG_10557 | ND | - |
| *UMAG_05317 | ND | - |

| | | |
|---|---|---|
| ***UMAG_05314** | ND | - |
| ***UMAG_05318** | g260, g261 | 38.99, 31.82 |
| ***UMAG_05319** | g262 | 37.65 |
| ***UMAG_05302**[d] | ND | - |
| ***UMAG_05303**[d] | ND | - |
| ***UMAG_10555**[d] | ND | - |
| ***UMAG_05308** | ND | 31.44 |
| *ros1*(UMAG_05853) | g445 | - |
| *cmu*(UMAG_05731) | g6240 | 63.33 |
| *hdp2*(UMAG_04928) | g4000 | 45.86 |
| *fox1*(UMAG_01253) | g4355 | 58.43 |
| *pit2*(UMAG_01375) | g5770 | 45.55 |
| *pit1*(UMAG_01374) | g5769 | 40.16 |
| *pit3*(UMAG_11316) | g5771 | 69.91 |
| *pit4*(UMAG_01377) | g5772 | 65.88 |
| *pep1*(UMA_GOX7E8) | g1278 | 48.12 |

ND indicates that no orthologs were detected at a cut-off of e-value 1e-5. Protein IDs in bold refers to the19A cluster, other identified effectors in lower case. From Brefort *et al.* 2014, a= Δ19A-1b, b= Δ19A-1a, c= Δ19A-1c, d= Δ19A-1d, * indicates right side of the 19A cluster.

We also looked for the presence of orthologs, in *P. graminicola*, to a set of well studied individual effectors (Table 2-9, lower case). Cmu1 is a novel type of hydrophobin, classified as a core effector withing 5 smut fungi (*U. maydis, S. scitamineum, S. reilianum, U. hordei* and *M. pennsylvanicum*) for which we found an ortholog in *P. graminicola.* P*ep1*, for which there is no recognizable domain, is crucial for the virulence in *U. maydis* and has a conserved function in *U. hordei* and *M. pennsylvanicum* (Lanver et al. 2017).

Pit2, inhibits cysteine proteases in *U. maydis* and other oomicete species (Lanver et al. 2017) displaying and unknown mechanism implicated in plant defense (Doehlemann et al. 2011). The Pit effectors family are localized in a cluster, also present in *P. graminicola*.

Hpd2 is a transcription factor in *U. maydis* which initiates the first wave of effectors (Lanver et al. 2017), whereas Fox1 is considered a late transcription effector, required for full virulence and host defence supression. Ros1 regulates the

developmental switch that leads to the formation of teliospores. Among its functions is to downregulate effectors, such as *cmu*, *pep1* and *pit2*, involved in supresing plant response (Lanver et al. 2017). Orthologues for these effectors in *P. graminicola* suggests the potential establishment of a biotrophic interaction with the host, as demonstrated for *U. maydis* (Lanver et al. 2010).

Furthermore, we found orthologues genes for secreted proteins from grass-infecting smuts (Schuster et al 2016). From the clusters 1428, 1431, 1425, 1311 and 1341, we found at least one orthologue for each cluster, except 1311 (Appendix. 2-). For the cluster 1464 (Appendix. 2-), which contains secreted Mig1 related proteins, we found four paralogs. In *U. maydis* the deletion of this family did not alter virulence whereas the deletion of orthologous in *S. reilianum* resulted in hypervirulence (Schirawski et al., 2010). It will be interesting to run the same experiment in *P. graminicola,* which results might help to elucidate the current proposition that these effectors may represent avirulence proteins (Basse et al., 2000).

Although we have not experimentally proven pathogenicity of *P. graminicola,* our findings suggest it has biotrophic features with a high possibility of infecting grass (or similar crops) by having the required machinery to trigger an infection process.

## 2.5 CONCLUSIONS

We sequenced and assembly the genome of *P. graminicola*, a smut fungi member and a MEL-C producer. To accomplished this we integrated two powerful platforms, PacBio genome sequencing and Illumina based RNA-seq. From this we produced a high quality assembled genome for *P. graminicola* CBS 10092, comprising a total of 6602 genes distribuited in 34 contigs. In addition, we accomplised to annotated approximately 84% of its proteome.

*P. graminicola* despite being classified within the *Pseudozyma* genus, its orthology shows more similarity to other genus such as *Ustilago* and *Sporisorium,* both comprising plant pathogens, meaning its classification needs some revision.

Additionaly, we identified orthologues for effectors from *U. maydis* and *S. reilianum* species in *P. graminicola's* genome, However, the number of orthologus to pathogenic genes in *P. graminicola* was lower than in these close relatives, suggesting a less virulent beviour and instead a biotrophic mechanism. Interestingly, no infection process has been described for this strain, meaning this species could be a potential plant pathogen.

This information is of importance to the fungal community as helps to streghten the current knowledge on the field and serves as a new available source of information for future comparative genomic and pathogenic studies, within smut fungi or even outside the *Pseudozyma* genus.

Alongside with the availability of the genome applicable to pathogenic studies, this new genome adds up to the very small database for smut fungi MEL producers, which currently comprises only 6 species. Noteworthy to mention *P. graminicola* genome assembly and annotatio contributes to the biotechnology field as a source of secondary metabolite production, as example with MEL, but definetely its 6603 genes must harbour more interesting proteins coding for metabolites which deserve to be studied.

In order to undertake an integrative analysis of the genome characterised in this chapter and direct this to understand the biology behind MEL production, in the next chapters we report the analysis of the MEL biosyntheic gene cluster, changes in the transcriptome associated with MEL production and investigate the roles of key transcription factors with potential regulation in MEL production.

CHAPTER 3

# 3 IDENTIFICATION OF MEL PRODUCTION BY *PSEUDOZYMA GRAMINICOLA* BY $^{1}$H-NMR

## 3.1 INTRODUCTION

### 3.1.1 Analytical identification of MELs

MELs are glycolipids comprising a hydrophilic head and a hydrophobic tail, the former being a sugar, mannosylerythritol, the latter a fatty acid (FA) (Figure 1-3) (Hewald et al. 2006; Konishi and Makino 2017; Morita et al. 2009a). MELs are arranged in four different configurations (A-D) based on acetylation and the FA side chains range from four up to 18 carbons (Jezierska et al 2018). Acetylation can occur at carbons four and six (MEL-A), singly, either at carbon four (MEL-B) or carbon six (MEL-C) or be completely absent (MEL-D) (Figure 1-3) (Morita et al. 2009a, 2013). MEL production varies with respect to the ratio of the four conformations, and this is species dependent (Moritaet al. 2007). Variation in the acetylation pattern and FA side chains results in a complex mixture with MELs with a range of structures being secreted (Hewald et al. 2006). For example, the major MEL-A producer *P. antarctica* T34 yields 70% of MEL-A, with the remaining 30% being both MEL-B and MEL-C (Kitamoto et al. 2001). *P. rugulosa* NBRC 10877 yields 68% MEL-A, 12% MEL-B and 20% MEL-C (Morita et al. 2006) whereas *P. tsukubaensis* JCM 10324T produces 100% MEL-B (Fukuoka et al. 2008). On the other hand, the species *Pseudozyma graminicola* has been reported to produce predominantly MEL-C (85%) and some traces of MEL-A.

Therefore, in order to identify different types of MELs in a heterogeneous mix the implementation of methods that allow discriminating 1) between configurations and 2) between carbon chain lengths is required. Among the most commonly used

methods are thin layer chromatography (TLC), mass spectrometry (MS) and nuclear magnetic resonance (NMR). Each of these techniques targets a different feature of the molecule, for example TLC usually focus on the hydrophobicity of the molecules, whereas MS detects the molecular weight (mass) and $^{1}$H NMR, the hydrogen profile of the molecules. MS and NMR are the most commonly used techniques in metabolomics, where the aim is that all or as much as possible of the metabolites are identified and quantified (Craig et al. 2006).

MS is an analytical technique that requires metabolites first to be ionised in order to acquire the mass-to-charge ratio (m/z) data. These resulting ionised compounds will provide a fingerprint of the original molecule by peak patterns. Due to the usual complexity of samples in MS metabolomics, a separation step is required prior analysis (Alonso, Marsal, and JuliÃ  2015), requiring of extra equipment such as columns for high-performance liquid chromatography (HPLC). The separation step is based on the interaction of the metabolite in the sample with the materials inside the column; therefore metabolites with different chemical properties will require different amounts of time to pass through the column. This time is known as retention time and coupled to the peak pattern provides an unique signature for the analysed metabolites (Alonso, Marsal, and JuliÃ  2015).

With reference to NMR, the most common type of analysis involves the use of hydrogen ($^{1}$H), as is the predominant isotope present in nature hence the most well studied (Marion 2013). The spectral parameters of hydrogen present in molecules are converted into angle and distances used to compute information about their structure (Marion 2013). $^{1}$H NMR has been extensively applied to the study of MELs (Faria et al. 2014; Kim et al. 1999; Konishi et al. 2010; Konishi and Makino 2017; Morita et al. 2006, 2008; Morita & Habe, et al. 2007; Saika et al. 2016; Sajna et al. 2013) for identification of the chemical structure and feature of the hydrogens conforming the FA and sugars from its molecule.

### 3.1.2 Fatty acid feedstocks for MEL production

The production of MELs has been reported using different FA feedstocks such as soybean oil and olive oil (Morita et al. 2009a; Morita, Konishi, et al. 2007). A very limited number of studies report MEL production using waste FA from sources such as refinery waste (Bednarski et al. 2004), soap stock and waste frying oil (Dzięgielewska and Adamczak 2013) or low-demand renewable sources like sunflower oil (S. et al. 1999). Consequently, although MELs have favourable chemical properties, their production remains expensive which limits their market potential.

The addition of FA to the feedstock makes the recovery and isolation of MELs a major challenge. This is due to the amphiphilic nature of the biosurfactants, requiring solvent extraction and/or precipitation of MELs prior to implementation of techniques such as TLC or HPLC to facilitate quantification and analysis. This increases the cost of manufacturing (Udo Rau et al., 2005) making the production process and recovery unattractive.

## 3.2 Chapter aims

I. To identify MEL production from *in house* and fermenter samples
II. To develop a $^{1}$H NMR protocol to semi-quantify MEL production
III. To evaluate different feeding and sampling methods over a 5 day time course fermentation under producing and non-producing conditions

### 3.2.1 Chapter description

This chapter describes the challenges and limitations associated with the identification and characterization of MELs produced by *P. graminicola* by TLC and NMR. We show the rationale step by step that led us to develop a standardised protocol for the detection and semi quantification of MELs on a background of

FA from flasks, a fermenter and micro-fermentations using a high-throughput biolector. We carried out comparisons between the three systems and tested different feeding patterns, sampling methods and fatty acid feed stocks, over a time course of 117 hours, in order to understand the MEL production in *P. graminicola*. In this respect, we developed and optimised and systemised protocol to detect and semi-quantify these biosurfactants without prior purification.

## 3.3 MATERIALS & METHODS

### 3.3.1 Media composition

Description of the media used on the different experiments performed detailed in tables below. All the media was autoclaved at 121°C at 15 lb of pressure. The glucose was filtered sterilised.

The growth media was used to replicate the cells to reach the exponential growth phase. Therefore, the concentration of glucose is higher than the producing media.

Table 3-10. Growth media composition

| Raw material | Concentration (g/l) |
|---|---|
| Glucose | 40.0 |
| Sodium nitrate ($NaNO_3$) | 3.0 |
| Mono potassium phosphate ($KH_2PO_4$) | 0.3 |
| Magnesium sulphate heptahydrate ($MgSO_4$ $7h_2o$) | 0.3 |
| Yeast extract | 1.0 |
| Deionised water | qs |

The producing media was used to induce MEL production based on recommendations from the industrial partner, CRODA. The concentration of sugar is half as this will be compensated by the addition of FA which will be explained next.

Table 3-11. Producing media composition

| Nutrients | Concentration (g/l) |
|---|---|
| Glucose | 20.0 |
| Sodium nitrate (NANO3) | 3.0 |
| Monopotassium phosphate (KH2PO4) | 0.3 |
| Magnesium sulphate heptahydrate ($MGSO_4$ $7H_2O$) | 0.3 |
| Yeast extract | 1.0 |
| Deionised water | qs |

The strain *Pseudozyma graminicola* CBS 10092 was obtained from the industrial partner CRODA and was grown on YM agar for 2 days at 30 °C.

### 3.3.2 Batch culture

Batch fermentations took place at the Institute of Integrative Biology (IIB) facilities using *P. graminicola* CBC10092 strain. An individual colony was inoculated into 50 mL of growth media (Table 3-10) at 30 °C and 200 rpm in an orbital incubator for 48 hours. 30 mL of this was used to inoculate 300 mL of producing media (Table 3-2) in a 1-L baffled flask. For MEL productions FA were added, whereas for the non-producing control flasks, FAs were omitted. The FA used was specific to CRODA. This has a composition of predominantly palmitic acid, stearic acid, oleic acid, linoleic acid and some erucic acid. This was named CRODAFAT. The optimisation process took place in two stages; in the first one we used batch cultures and in the second one automated micro-fermentations.

#### 3.3.2.1 Batch feeding

In batch, we implemented two feeding systems; "*single dose*" in which a total of 16 g of FA (per 300 mL of culture) was introduced at the start of the fermentation without further additions. The second, "*multiple dose*" involved the addition of 3.84 g of FA every 24 h. Samples of 50 mL were taken every 24 h during a five day time course. Each sample was centrifuged (10 min, 10000g, at room

temperature) and three layers were observed in the producing samples (addition of FA) whereas only two layers for the non-producing samples (no addition of FA) (Figure 3-1). Each phase was transferred into a plastic container suitable for its volume and stored at -80°C. We worked with three replicates per time point (24, 48, 72, 96 and 117 h) for each condition (producing and non-producing). Diagram depicting methodology in Figure 3-2.

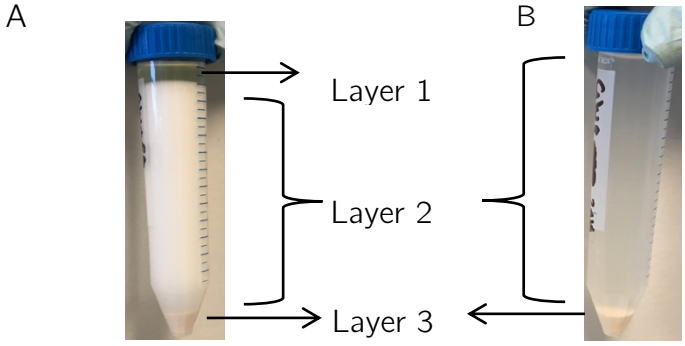



Figure 3-1. Indication of layers present on centrifuged samples under the presence and absence of FA. Culture were grown in media containing glucose and FAs (A) or only glucose (B) as carbon source. Layer 1: FA. Layer 2: Media (plus secreted MELs in samples when FA feedstock provided). Layer 3: Cell pellet.

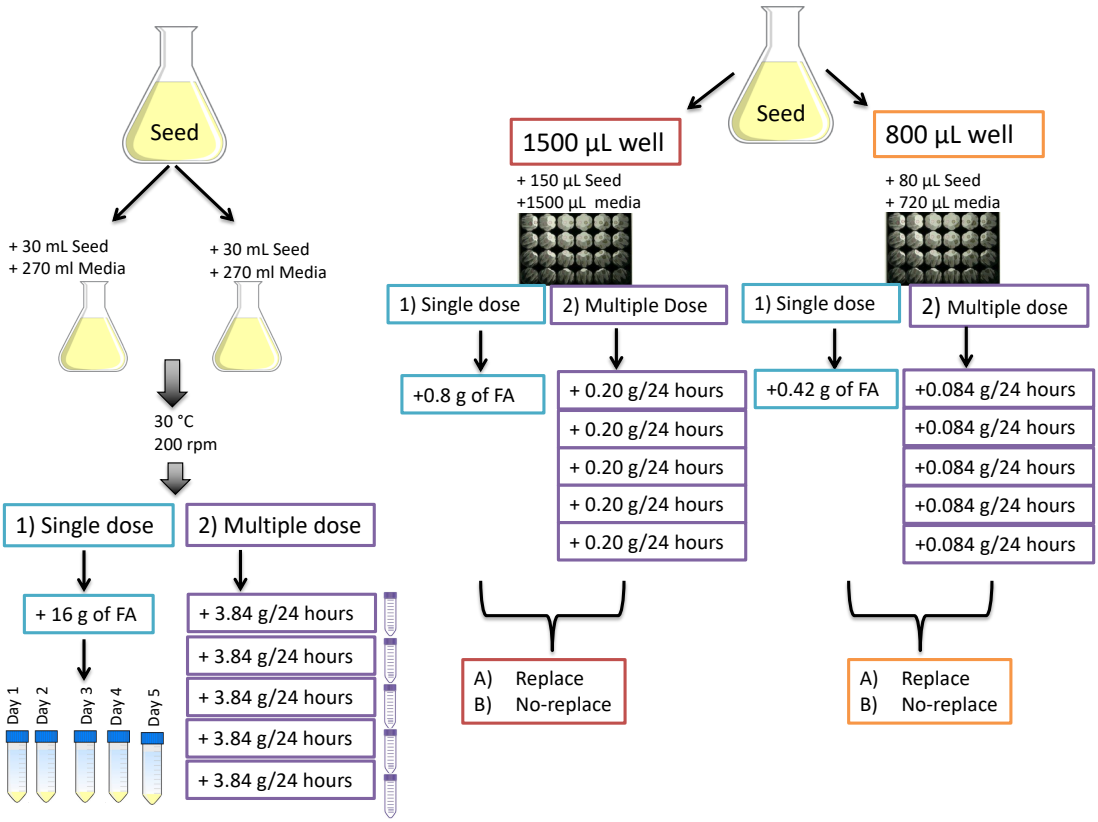



Figure 3-2. Diagram for MEL production standardisation on (batch) left panel and microfermentation (right panel) system.

### 3.3.3 Micro fermentation culture and feeding

Micro fermentations were conducted using a high throughput robotic fermenter Biolector XL (m2p labs, Baesweiler, Germany; at 30 °C; 1,200 rpm) where we tested two media volumes for the feeding systems, 800 µL and 1500 µL. As in the batch phase, a 48 h seed culture was used to inoculate the media at a ratio of 1:10. For the "*multiple dose*" method the dosage of FA varied according to the final volume of the well; 0.05 g for the 800 µL culture, 0.096 grams for the 1500 µL culture, every 24 hours. A sample of 200 µL was taken at each time point and the total volume was used for the NMR analysis (method described in next section). We implemented two sampling systems; a "closed system" where the volume decreased with each sample taken (no replacement of media) and a "continuous system" where the fresh media was added to replace the sample volume. The set up for this experiment is depicted in Figure 3-2 and Figure 3-3.

These cultures were grown for four days at 30 °C. These fermentations were monitored using the Biolector coupled to the Robolector XL.

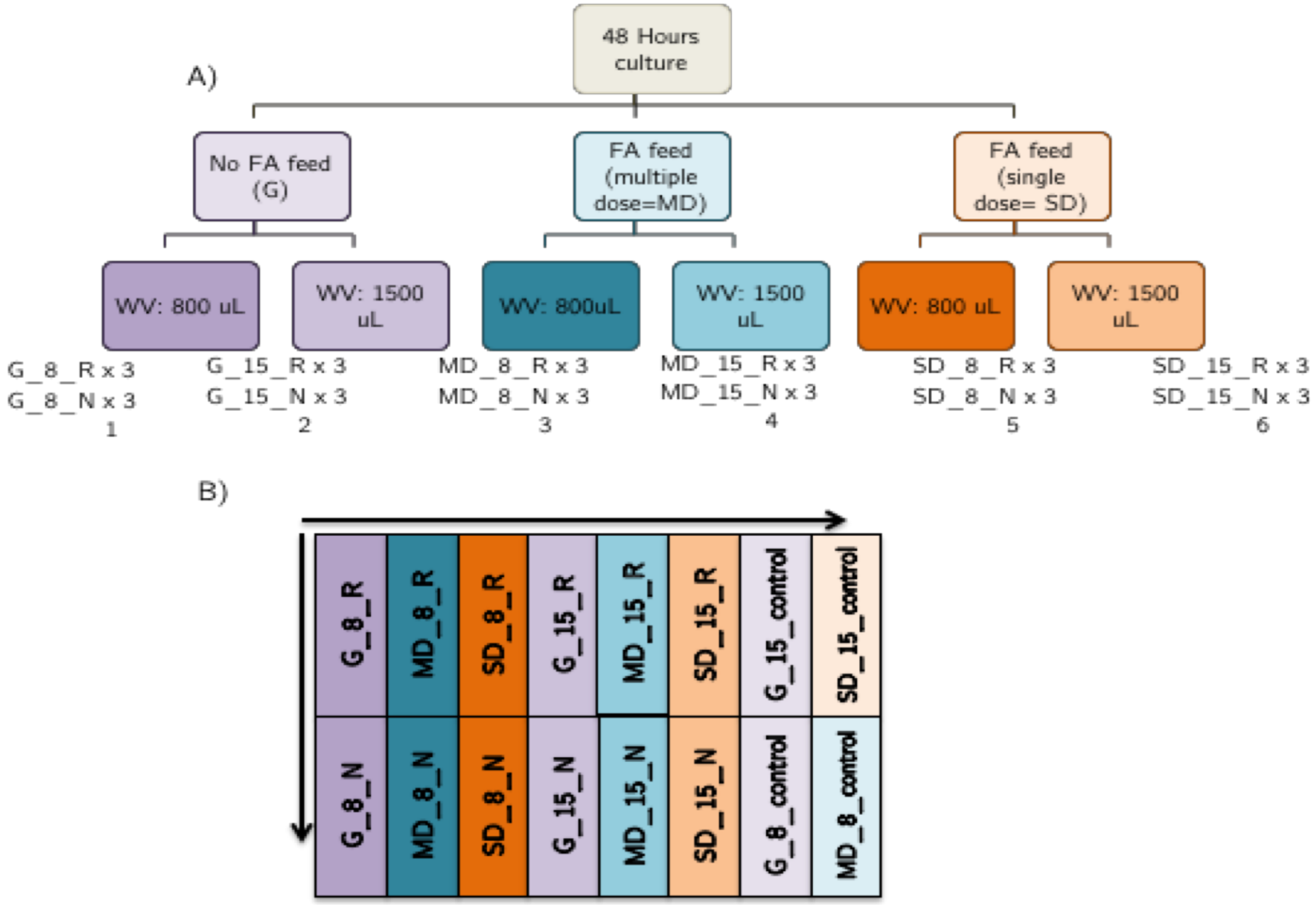



Figure 3-3. Microfermenter experimental design. A) Diagram of sampling set up using the Biolector for parallel fermentations. Each condition (designed with a number) was carried out in triplicate. WV: total well volume. B) 48 well format shown to illustrate order in which samples are collected, each fermenter condition is given a code. Cultures were either uninduced with no FA (U) or FA dosed every 24h (MD) or once until the end point (SD). Two well volumes were used, either 800 µL(8) or 1500 µL(15). Final letter stands for type of feeding system, either where the sample volume was replaced with fresh media (R) or not replaced (N).

### 3.3.4 MEL standards

MEL standards were provided by CRODA. These were purified through an extraction process at their installations. The standards were produced by their industrial strain *P. aphidis*. Due to the extraction process (which was kept confidential), and the similar properties of MELs, these standards were given as: mixed fractions enriched with MEL-A, MEL-B/MEL-C, or MEL-D and a crude fraction, which corresponds to a mix of all the fractions plus potential traces of CRODAFAT.

### 3.3.5 Thin Layer Chromatography (TLC)

We implemented the anthrone method as reported by Morita and collaborators (2006) with some variations, which included a dilution of the anthrone mixture (1 g of anthrone in 500 mL of 72% sulfuric acid) at a 1:1 ratio with pure ethanol.

### 3.3.6 Spectral processing for detection and quantification of MELs by $^{1}$H Nuclear Magnetic Resonance (NMR)

The FA layer was lyophilised overnight and resuspended in 600 µL of deuterated chloroform (to label hydrogens), vortexed and spun down at maximum speed for five min at 10 °C. The supernatant from this mixture was transferred to 5 mm, outer diameter, NMR tubes and analysed in a 600 MHz Avance III spectrometer (Bruker) equipped with a R1 cryoprobe X autosample (sample Jet) using cpmgpr1d filters for small molecules via a Carr–Purcell–Meiboom–Gill (CPMG), at the NMR Centre, University of Liverpool.

Each spectrum was individually processed using the software TopSpin 3.5 patch level 6 (Bruker®) for quality control (QC) by checking the width of the chloroform peak at 7.26ppm. Based on other chloroform extractions run at the NMR facility a line width at half height of approximately 1.1 Hz was used as optimal. High quality spectra must have a single peak at the chloroform position,

with a smooth curve and flat uniform baseline (absence of size wave) (Figure 3-4). Using the spectra from this project a standard deviation of 0.4 Hz was obtained. In this respect, a range of 0.7 Hz to 1.5 Hz was set as acceptable. Any spectra that did not pass quality control (determined by a poor line width of the chloroform peak) were centrifuged further prior to analysis, as this was indicative of sample heterogeneity, which was typically due to precipitant or poor phase separation (aqueous solution present).

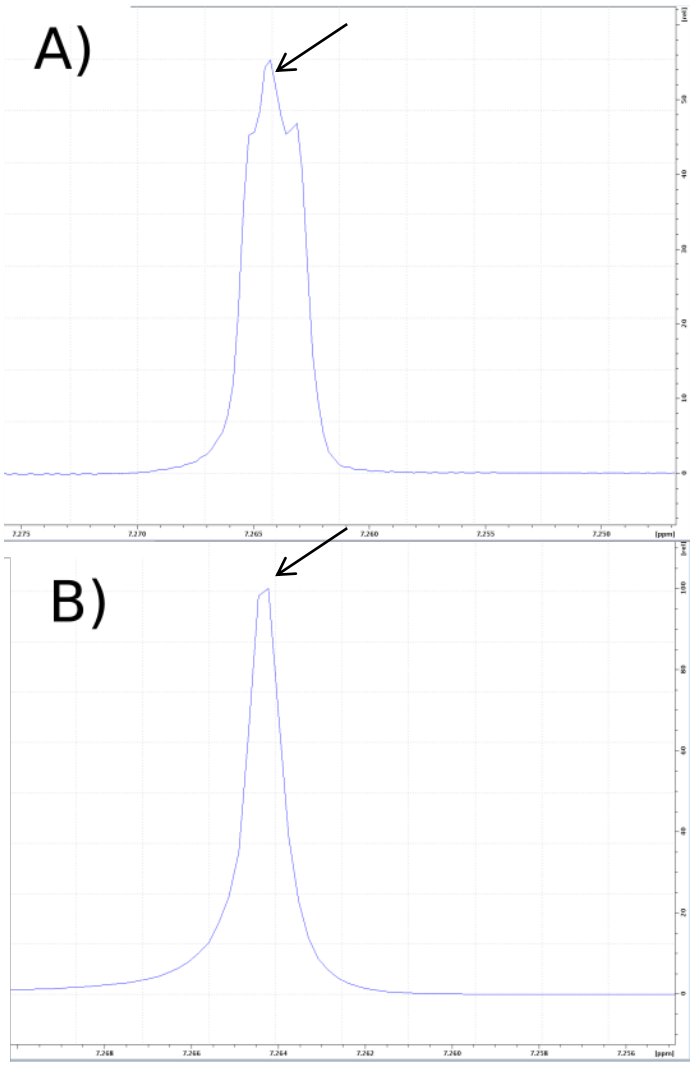



Figure 3-4. $^{1}$H-NMR QC spectra examples for chloroform peak at 7.26 ppm. A) Low quality spectra. B) Optimal QC value, included in data set.

The spectra that fulfilled the QC values were used to prepare the pattern file, which is a decomposition of each peak from all the aligned spectra. Each peak was divided into regions that correspond to different components/molecules from the media, including FA and mannose (Figure 3-5). Once the peaks were identified the corresponding positions were annotated and used to create the bucket table (Appendix 3-1).

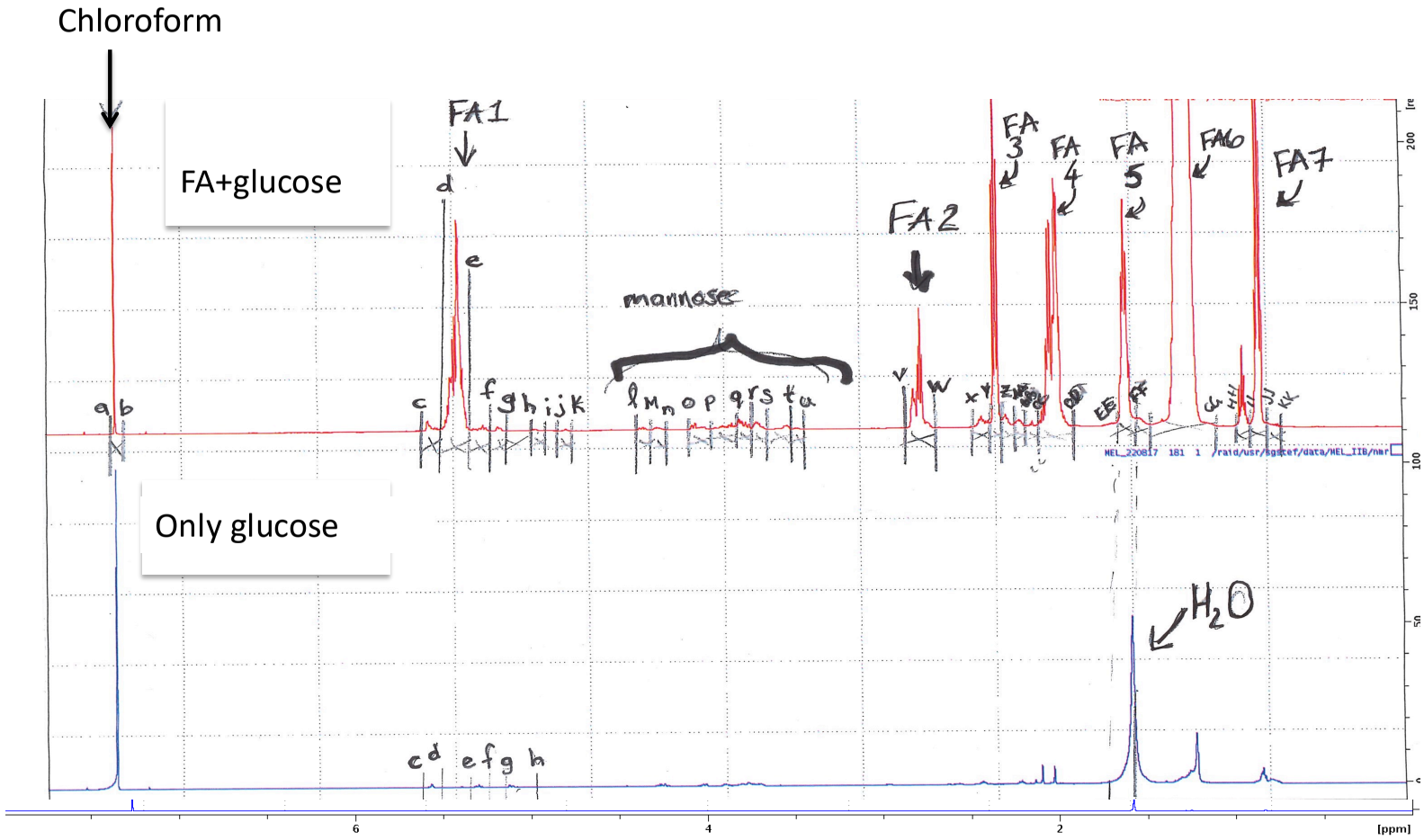



Figure 3-5. Spectra from fermentation in the presence of FA and glucose used to create pattern file. An NMR spectra is presented that has been manually annotated with letters corresponding to the decomposition of the spectra. Upper panel (red) spectra of producing sample (FA+glucose), lower panel (blue) spectra of a non-producing sample (no FA). Spectral analysis and statistics

The spectra was obtained by collecting the intensity of the labelled hydrogen with the deuterated chloroform. The spectra were integrated with Amix (Bruker GmbH, Karlsruhe, Germany) in order to obtain a bucket table or spreadsheet with the corresponding intensities for each peak in the spectra, from the regions in the pattern file. Due to the nature of the NMR technique, the values reported after processing the spectra with AMIX represent a relative measure of the concentration of the metabolite; this software hence allows the transformation of intensity values to concentration values assigned from the labelled hydrogen. Data was collected for all the experiments

The data was normalised using the deuterated chloroform peak as the reference (7.26ppm) (Craig et al. 2006), with MEL concentration expressed as a proportion of the solvent. These normalised data were used to identify differences in the quantity of MEL produced. This normalisation was calculated using an R script developed by the Computational Biology Facility at the University of Liverpool.

Data transformation tool place by taking the cube root of the data values in order to make the scale contiguous and when looking for comparisons between samples Pareto scaling was applied, which involves a mean-centering and division by the square root of standard deviation of each variable. After this we semi-quantified the production of mannose-related and MEL-related metabolites between experiments (conditions, time points, type of FAs) by a one-way ANOVA and applied a Fisher test, at a significant value set at 95% of confidence. Principal Component Analysis (PCA) was used to identify differences between replicates. The statistics were calculated using MetaboAnalyst online platform (Chong et al. 2018) and R to plot graphs using ggplot2 package (Wickham 2016).

## 3.4 RESULTS & DISCUSSION

### 3.4.1 Limitations of MEL identification

We used the media layer from batch cultures (Figure 3-1) of *P. graminicola* to identify the presence of MELs. Our first approach to detect MELs was by TLC, as this is considered to be a quick, simple and economic technique.

Based on literature review the anthrone-sulphuric method appeared to be the most widely used for MEL identification (Flagfeldt et al. 2009; Konishi and Makino 2017; Morita et al. 2009b; Sajna et al. 2013; Yoshida et al. 2014). Its basis is the reaction of the carbohydrates, under the acidic conditions, with anthrone dissolved in concentrated sulphuric acid, creating brown-green spots. By this method we were able to detect the presence of MELs. However, the quality of our TLC plates was suboptimal (Figure 3-6) as the separation of the different compounds wasn't clear, with the carbohydrates appearing smeared. Nevertheless the spots corresponding to the MEL-C, as defined by the standards, fits with *P. graminicola* MEL reports (Morita et al. 2008). In addition, the CRODA standards showed low purity (Figure 3-, lines 2-3,5-6), as multiple compounds were observed; this might be due to an unsatisfactory purification of the standards during the extraction process and/or subsequent degradation products.

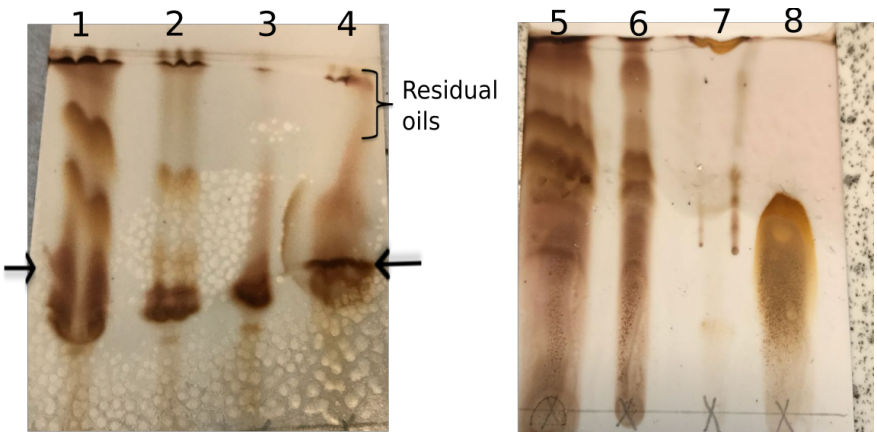



Figure 3-6. TLC produced with CRODAFAT by *P. graminicola* and displayed of MEL standards. MEL-C, produced by *P. graminicola* (Line 4), from media of a 72 hours fermentation with CRODAFAT as FA source, compared to different versions of MEL standards. The MEL standards were provided by CRODA as 1) a mixture of the four versions of MEL, enriched with: 2) MEL-A, 3) MEL-D; a sample of 5) not purified crude mix, 6) diluted not purified crude mix, 7) Commercial olive oil and 8) CRODAFAT. Therefore, the integration of all the bands allowed us by discrimination, the identification between MEL types.

#### 3.4.1.1 $^{1}$H NMR analysis

The $^{1}$H-NMR was conducted at the NMR Centre for Structural Biology (Liverpool). Our rationale to identify MELs in a high FA background was based on the amphiphilic nature of the MEL molecule, where once lyophilised and resuspended in deuterated chloroform, only the hydrophobic molecules will remain in the mixture, consequently only the mannose attached to a FA chain (Figure 1-3) will be visible in the spectra.

From the spectra, at the region where sugars are located (5-3 ppm) we were able to identify the presence of mannose, although discrimination between MELs variants among the standards was not possible (Figure 3-7) due to impurity and complexity of the mixtures.

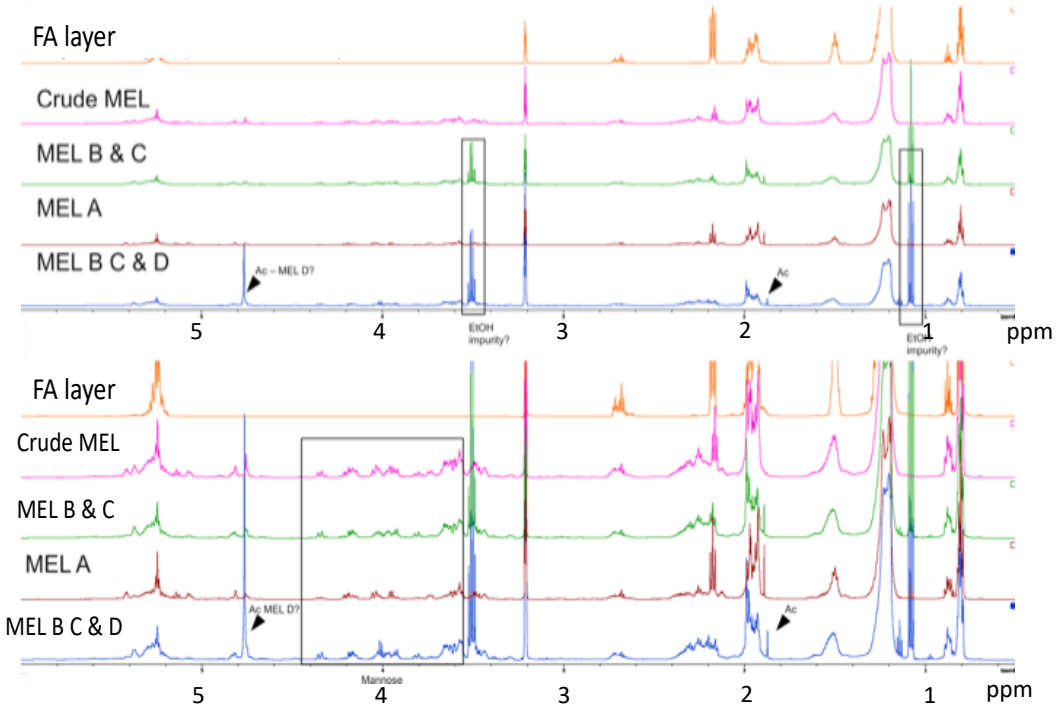



Figure 3-7. NMR analysis of MEL standards and FA layer. The MEL region of the $^{1}$H CPMG spectrum of different MEL standards (MEL-A, MEL-B/MEL-C, MEL-B/MEL-C/MEL-D and Crude) and FA layer at 600 MHz are presented. Upper panel displays region 5-0ppm, the lower panel shows 2X magnification of the spectrum scale. Mannose and possible ethanol impurity are highlighted in black boxes. Putative acetylation marked with arrows.

Analysis of the region 4.4-3.5 ppm showed a complete absence of mannose in the FA layer from our pilot fermentation (Figure 3-1, layer 1), suggesting any secreted MELs would be found as a mixture with the media (Figure 3-1, layer 2)

We also analysed cell extracts from the fungal pellet obtained by centrifugation (Figure 3-1). These extracts required significant optimisation as the presence of FAs and the intrinsic nature of the yeast cell (presence of a membrane and a cell wall) negatively affected the analysis. We assessed eight different extraction methods (Appendix 3-2) and those that gave the best results were method 1, 6 and 7 in terms of reproducibility and spectral quality. However, the presence of MELs in the culture media, the need for further protocols optimisation for analysis of the intracellular metabolome coupled to its increased complexity and the need

of an extensive fungal database, which was not available at the NMR Centre, motivated us to focus our work to only the extracellular layer.

$^{1}$H NMR was our chosen method for optimisation to facilitate MEL identification and semi quantification using media samples. From the region 4.4-3.5 ppm we were able to distinguish MEL-A from MEL-B/MEL-C in the crude mixture (Figure 3-), but not MEL-D. This crude mix corresponds to the four MEL standards given by CRODA plus some potential traces of fats and other compounds from their production processes.

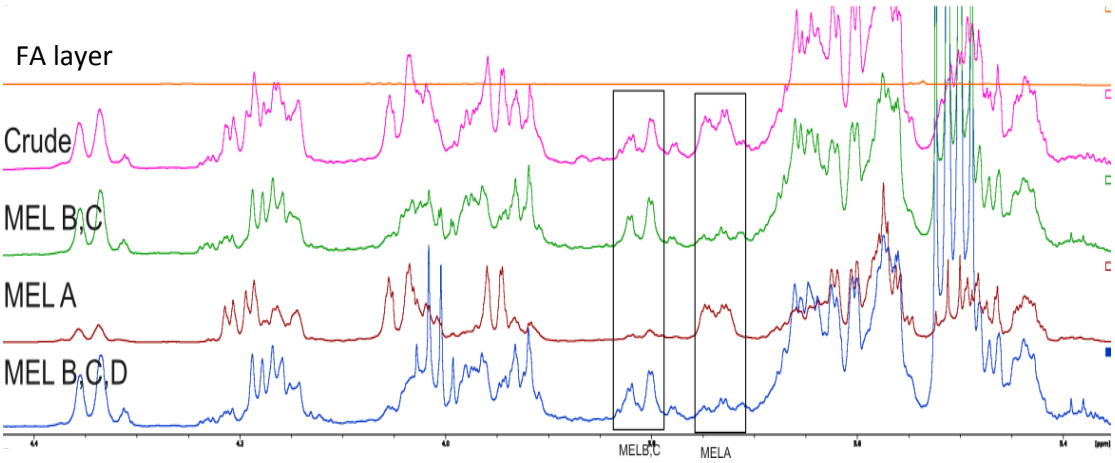



Figure 3-8. NMR analysis of mannose region from MEL standards distinguishing MEL B/C from MEL-A. The mannose region of the $^{1}$H CPMG spectra is presented. The crude mixture of the four versions of MEL standards, its enriched fraction and the FA layer were analysed at 600 MHz. MEL B/C and MEL A related signals are indicated (boxes).

The vast majority of reported $^{1}$H NMR analysis of MELs were conducted at 400 MHz (Konishi et al. 2010; Morita et al. 2006, 2008; Morita, Konishi, et al. 2007; Saika et al. 2016), a few at 500 MHz (Faria et al. 2014; Sajna et al. 2013) and only one at higher resolution, such as 600 MHz (Konishi and Makino 2017). These differences make comparisons inaccurate, being at 400 MHz the closest report for *P. graminicola* CBS 10092 (Morita et al. 2008).

### 3.4.2 Lyophilisation & Resuspension of samples in Deuterated Chloroform ($CDCl_3$) from batch cultures

As discussed above, in our case $^1H$ NMR analysis of the culture media was the most suitable technique to identify MELs. We next optimised the detection of MELs, which is made difficult due to the high background level of FA arising due to the growth regime (see section 3.3.2). According to the only published study for MEL production by *P. graminicola* (Morita et al. 2008), a culture including a total of 24 mL of soybean oil, grown over seven days, the total yield of MEL-C reported was 8.16 g/L. Based on this we expected the concentration of MELs to be low relative to the total FA content of the culture.

### 3.4.3 NMR spectral assignment for MELs from media layer under two different feeding systems

Based on a review of the literature, we determined two methods to feed the cells with FAs, one in which the FA was added from the beginning with no subsequent dosing (*single dose –SD-*), the other where a dose of FA was added every 24 hours over the full time course (*multiple dose –MD-*). It is worth noting that our industrial partner practices the second method for production of MELs.

We implemented both feeding systems with wild type *P. graminicola*, using 300 mL batch cultures and 1500 µL micro-fermentations, (section 3.3.2) and obtained a semi quantitative assessment of MEL production. From the NMR spectra, we observed mannose associated signals build over time for both of the treatments in which fatty acids were added; these signals were absent in the control (no FA) (Figure 3-18). It is noteworthy that the glucose concentration is about 6 times higher than nitrogen, and despite the latter being present in the initial media, it is likely to be exhausted over the 96 hours of the fermentation. Furthermore, it is unlikely for glucose to be present in the spectra, acquired in $CHCl_3$, as a polar molecule it would not be soluble in the hydrophobic solvent.

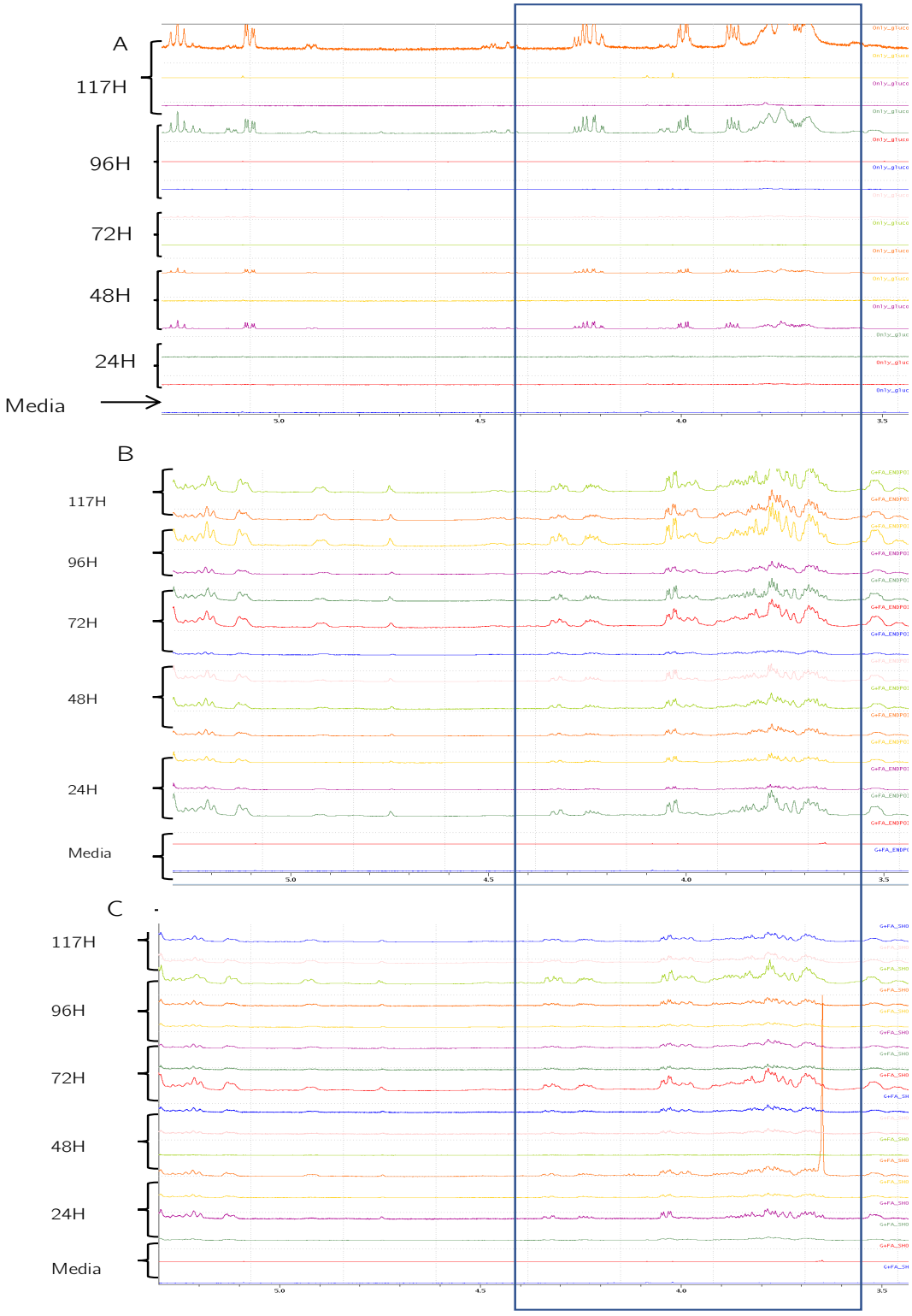



Figure 3-10. Sugar built up over time produced by *P. graminicola* flask fermentation under producing and non-producing conditions. Mannose region of the $^{1}$H CPMG spectra of *P. graminicola* media flasks fermentations at 600 MHz in deuterated chloroform. N=3. A) Control fermentation (n=2 for 24 h and 72 h), B) Single-dose regimen (n=2 for 96 h), C) Multiple-dose regimen (n=3). Mannose signals between feeding regimes (lower in control flasks) marked with black box.

In addition, we observed in absence of FA (control flasks) the highest MEL-related metabolites occurred at 72 hours (Appendix 3-2), suggesting MEL production might be linked to growth or development.

We observed the highest production of MEL-related compounds under the *single-dose* regimen; when compared to the *multiple-dose* regimen (Figure 3-11), having the mannose-related compounds the same trend (Appendix 3-). When FAs were added to the media, a gradual increase in mannose-related metabolites was observed (Appendix 3-), whereas when FAs were dosed, this increase had an erratic behaviour probably attributed to the difficulty in pipetting accurately small volumes of viscous heterogeneous material or/and due to a potential exhaustion of the FA, provoking a decrease in the production until the new addition (Figure 3-11).

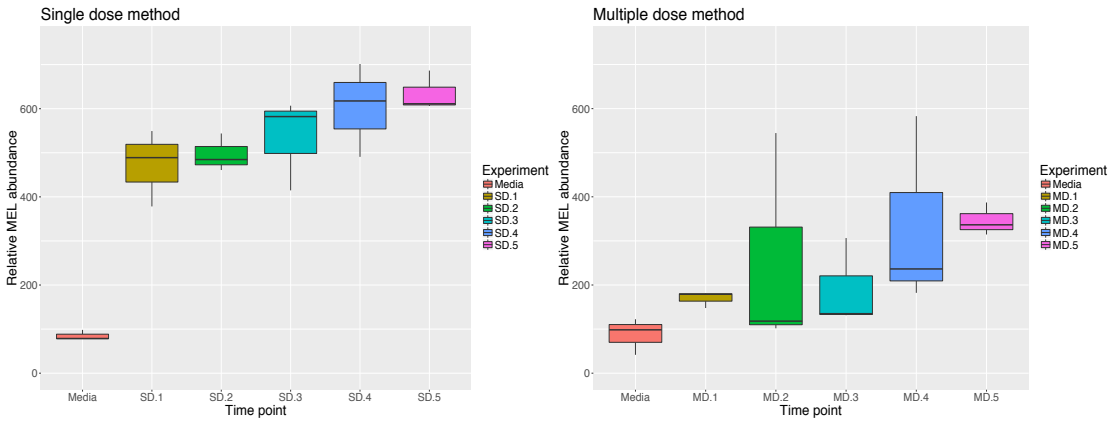



Figure 3-11. Comparison of MEL production in flask fermentation by *P. graminicols*. Relative abundance box plots of normalised MEL-related metabolites over a 117 time course flask fermentation under two different feeding regimes. MD= multiple dose regimen. SD= Single dose regimen. 1=24 h, 2=48 h, 3= 72 h, 4= 96 h, 5= 117h. N=3.

In order to evaluate the correlation between replicates we plot PCA to test the reproducibility of our method, for this we made use of one of the advantages of the NMR technique which is its sensitivity and the direct correlation existing between the signal from the spectra and the concentration of its corresponding metabolite (Alonso et al., 2015). This reveals high variation between replicates,

as demonstrated by principal component analysis (Figure 3-) and this is more evident as the concentration of FAs increased (Figure 3-B,C).

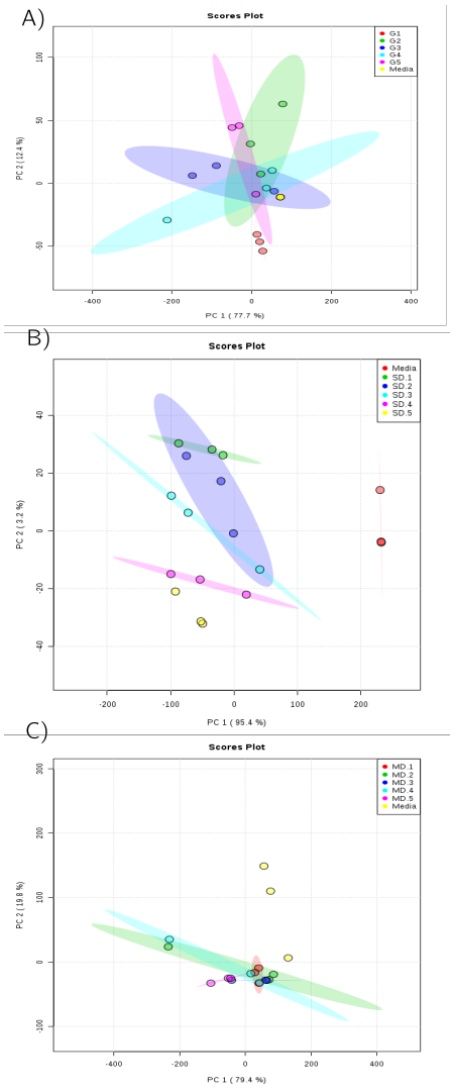



Figure 3-11. Principal component analysis from flask fermentations by *P. graminicola* showing variation between replicates. A) Control (no FA), B) Single dose method (SD), C) Multiple dose method (MD). N=3. Numbers on each condition refer to sampling time points: 1= 24 h, 2= 48 h, 3= 72 h, 4= 96 h and 5 = 117 h.

From the optimisation process we developed using *P. graminicola* CBS 10092, we identified some drawbacks on the set-up of the experiments. In NMR, normalisation and scaling of spectra is required; the former to make samples directly comparable to each other by reducing its variance, the latter typically to get parsimonious PCAs (Craig et al. 2006).

The normalisation is generally done to an internal standard. In our case major differences probably arise during pipetting due to the viscous nature of the media and FA leading to drastic differences between samples and replicates. Therefore, we used chloroform as internal standard for normalisation; this gave us the best results (data not shown). We also identified that the presence of FAs significantly lowered the QC values of the spectra. To overcome this problem, we implemented both, chloroform and area under the curve normalisations, which allowed us to optimise and systemise the data analysis.

### 3.4.4 Automated fermentation platform to develop a robust analysis of MEL levels in media

As our goal was to develop a reliable standardised methodology that enables the detection of MELs in media, we aimed to optimise the sampling method by implementing an automated system. We utilised a Robolector, a robotic automated fermentation platform that allows parallel high-throughput cultures in plates of 48 wells, with the capacity of precisely and accurately handle liquids (to sample and to dispense as required). This platform is coupled to a Biolector that monitors and records, in real time, physical parameters such as pH, temperature, dissolve oxygen and optic density as a measure of biomass.

In this manner, we aimed to obtain more consistency between replicates during sampling and consequently less variability. Previously we determined that the best FA feeding method for flask system consisted of the addition of a FA in a single dose. We therefore aimed to determine if the same principle applied to microvolumes. Additionally, we tested different well volumes (800 µL and 1500 µL) and different sampling methods: *continuous*, where the 200 µL samples was taken and replaced with media, and *closed* where the 200 µL sample was not replaced. The rationale of this was to compare an open industrial fermenter system (where replacement of media takes place) and a closed-flask system (where there

is no replacement)[1]. Testing of this kind with *P. graminicola* or a related species has not been reported in literature before and may potentially provide insights to improve set-ups in automated systems were the goal is to improve the repeatability for future protocol development. As expected this automated sampling showed overall more consistency between replicates (Figure 3-) and regardless of the sampling method or well volume, the MD regimen gave the most uniform data (Figure 3-C).

[1] See Section 3.3.3. and Figure 3-

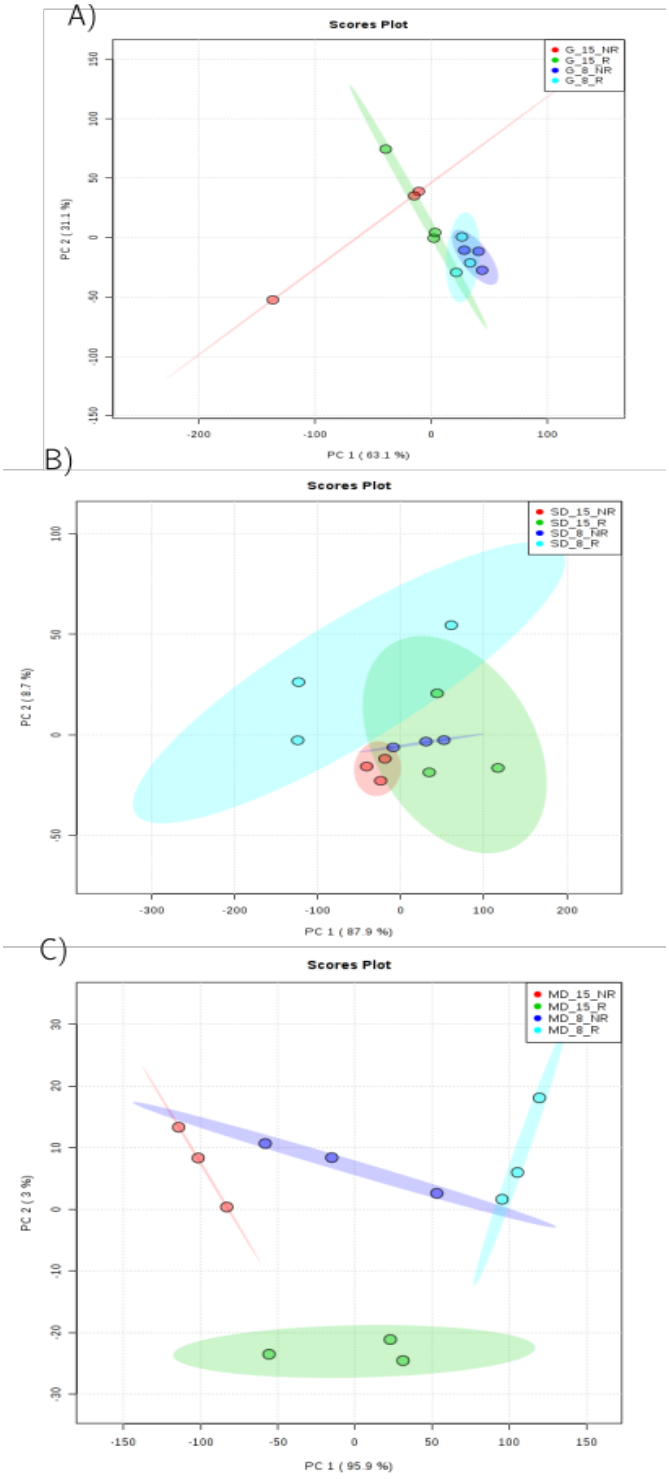



Figure 3-12. Principal component analysis for biolector end-point fermentations. 72 hours fermentations of *P. graminicola* were conducted using different well volumes and sampling systems. The resulting data was subjected to principal component analysis to assess variation between replicates. A) Control method (-G-, no FA), B) Single dose method (SD), C) Multiple dose method (MD). N=3. First letter stands for type of media, G: glucose, S: shot, end-point. Number stands for final well volume= 8: 800 µL, 15: 1500 µL. Final letter stands for type of feeding system, R: replace, N: no replace (i.e. G_8_R: non-induced system 800 µL well with replace of media).

We also observed that supplementation with FAs considerably decreased repeatability (Figure 3-) and growth rates were higher for the 1500 µL well volume, when using a closed sampling system under the MD regime (Appendix 3-5 $p < 2e\text{-}16$). Increased readings were more dramatic when under the MD regime; likely attributable to an interference on optic density readings due to an excess of FA.

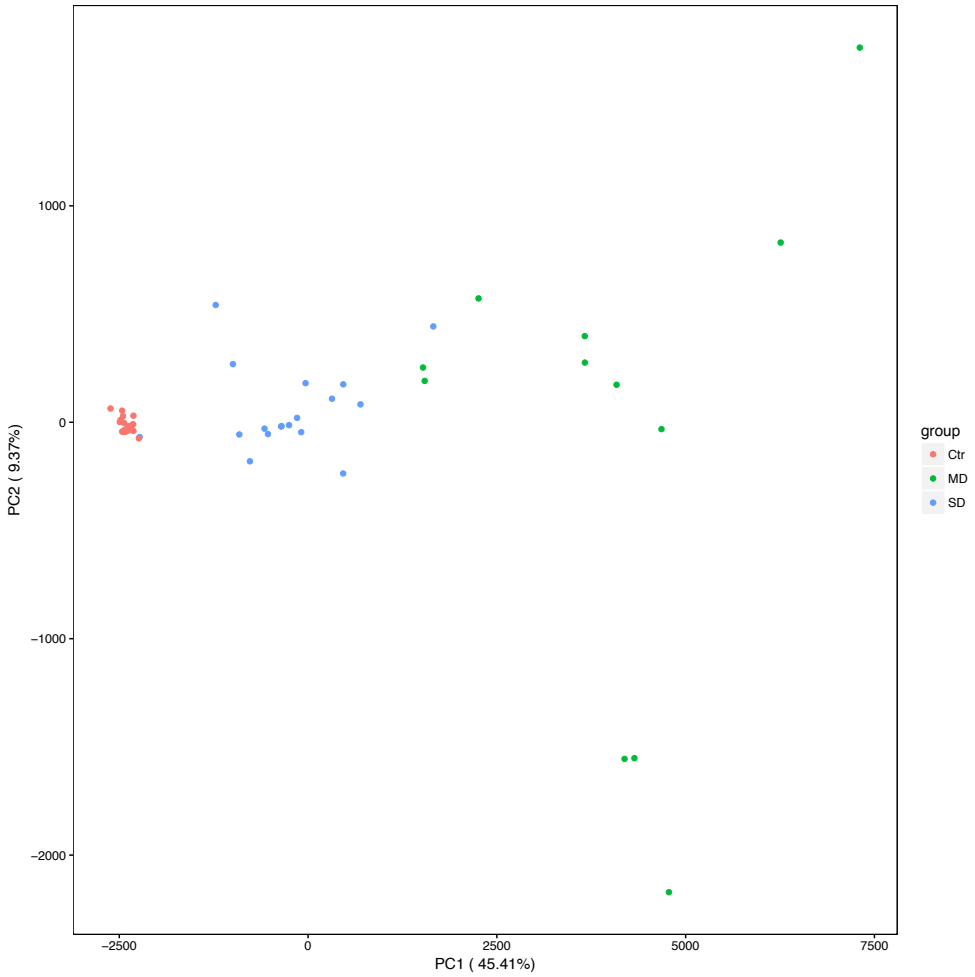



Figure 3-13 Principal component analysis for growth rates in micro-fermentations for *P. graminicola*. Cellular growth for control samples (Ctrl, no FA) and two feeding systems: multiple dose (MD) and single dose (SD). Highest growth rate variation for MD samples. Detailed key for each group (Ctr, MD and SD) in Appendix 3-.

Based on the preliminary analysis, we increased the number of replicates to six, to help compensate for the high variability of FA supplemented samples and included media controls with and without FA. We used a total well volume of 1500 µL with

no replacement of media during sampling and a feeding system of dosing every 24 hours during four days. With this optimised regime we observed an improvement in sample repeatability between replicates, based on principal component analysis (Figure 3-) and growth rates (data not shown).

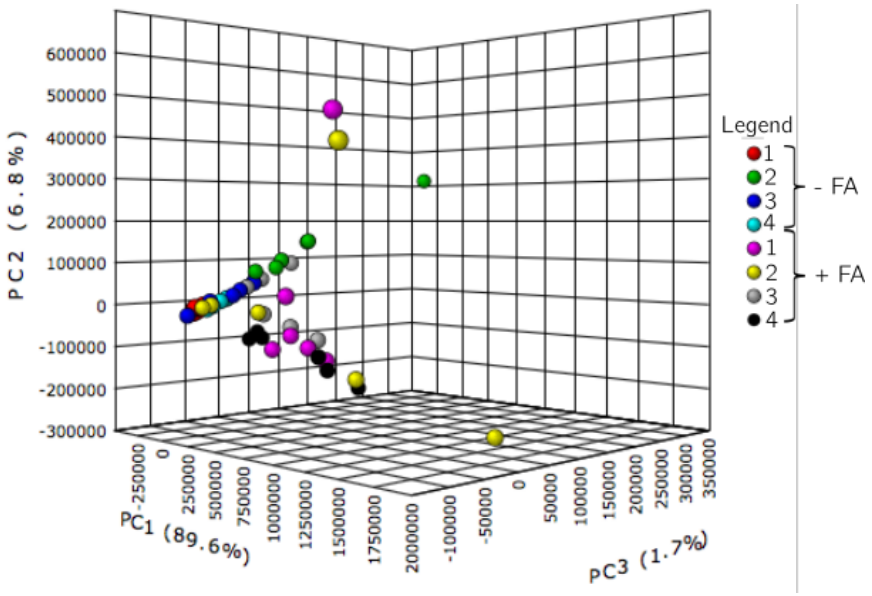



Figure 3-14. Principal component analysis for time-course micro fermentations by *P. graminicola*. N= 6. Numbers on each condition refer to sampling time points: 1= 24 h, 2= 48 h, 3= 72 h and 4= 96 h. Samples from wells in absence (-FA) or presence (+FA) of fatty acids

Our method showed very low or no evidence of MEL-related metabolites (lipophilic sugar moiety) in cultures to which no FA was added. Additionally, at 48 hours we observed the highest concentration of mannose-related compounds, which subsequently decrease over time. For FA supplemented cultures MELs were observed at significant higher concentrations at all time points, when compare to non-producing conditions (Figure 3-12). Additionally, there was less variation between time points (Figure 3-).

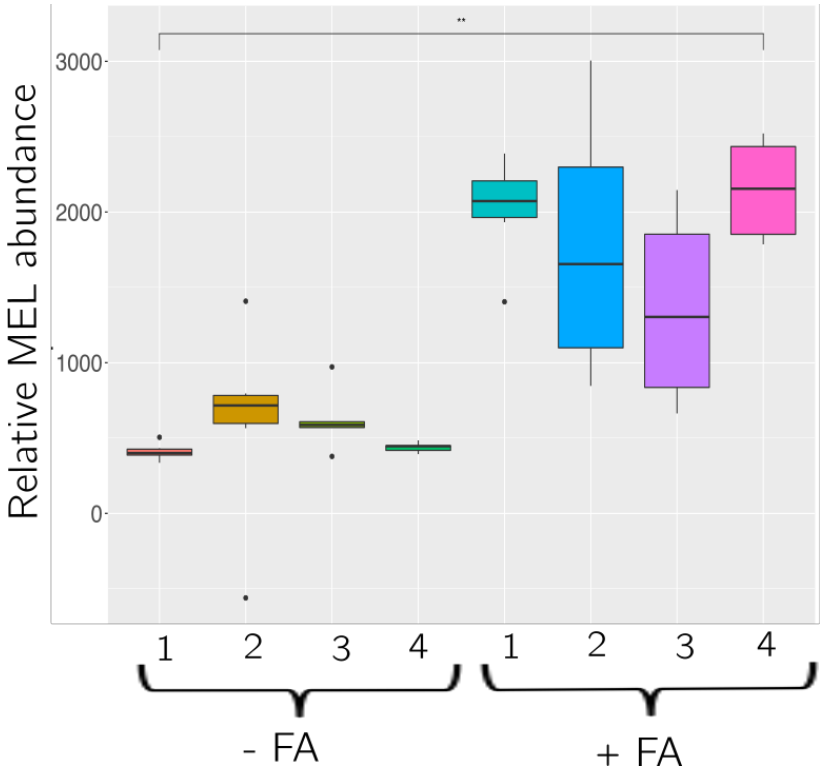



Figure 3-12. MEL production comparison over a 96 hour time course micro-fermentation. Micro fermentation was conducted in absence (-FA) and presence (+FA) of FA for *P. graminicola*. Relative abundance box plots of normalised MEL-related metabolite. N=6. Numbers on each condition refer to sampling time points: 1= 24 h, 2= 48 h, 3= 72 h and 4= 96 h. Comparison between groups statistically significant, **: p <= 0.01.

### 3.4.5 *P. graminicola* WT under different fermentative conditions

Unlike *P. aphidis* DSM 70725 (Rau et al. 2005), in which MEL production starts earlier in bioreactor when compared to flasks in *P. graminicola's,* MEL production overall has similar timing regardless the fermentative system: flasks (Figure 3-11) or micro-fermentations (Figure 3-12). Despite production increasing after 24 hours regardless of the system, the yields are remarkable different, therefore factors such as feeding system (dosing or fixed concentration of the FA) are more important for the production dynamics.

In summary, we have monitored the MEL production by *P. graminicola* during a 96 hour time course using glucose and waste FA as carbon sources. We identified the drawbacks of setting-up flasks experiments and we optimised a protocol to detect MELs even at low micro-volumes (800 µL).

### 3.4.6 Flask vs 96 well format (Robolector)

MEL production in basiomycetes has been extensively studied (Faria et al. 2014; Hewald et al. 2006; Konishi and Makino 2017; Morita et al. 2009a, 2013; Morita, Konishi, et al. 2007; Rau et al. 2005, 2005) and yet its standardisation with respect to a variety of factors has not been reported. For different strains the FA dose rate have been shown to be critical. Rau and collaborators (2005a) found a difference of 20 g/L in MEL yield between two *P. aphidis* strains (DSM 70725 and DSM 14930) when using soybean oil at the same dose rate. They also found a difference of 75 g/L of total MEL produced by the same strain (*P. aphidis* DSM 14930) when "uncontrolled feed substrates (glucose, nitrate, yeast extract) were added after nitrate limitation". Konishi and collaborators (2008) also found remarkable differences in MEL yield when glucose, FA and yeast extract were added to fed-batch fermentations with *P. hubeiensis* KM-59. In addition, Rau and colleagues (2005) identified differences in MEL yield among strains of *P. hubeiensis*.

We, noticed differences in the MEL production by *P. graminicola* according to the time point, sampling (media replacement, non-replacement) and FA supplementation regimes. In the flasks system we noticed green, bead like, cell aggregates at the bottom of the flasks after three days of fermentation (data not shown), an observation that was also reported by (Rau et al., 2005) and identified as an indicator of enhanced MEL production. This implies a link with cell morphology and development and possibly intercellular interactions to MEL biosynthesis.

We also identified differences between dosing regimes (MD, SD), being higher than 3 fold between systems (flask and microvolumes, Figure 3-11, Figure 3-12 respectively) highlighting the differences that were also observed in the behaviour of production, as this did not follow a gradual increase, demonstrating micro-fermentations as the best system. As a final result we provided with an standardised method which is depicted in Appendix 3-7.

### 3.4.7 Robolector multiple-dose vs Robolector single-dose

The feeding system did not have a significant effect in MEL production. Even when we observed variations on MEL-related metabolite, according to how the dose occurred. A FA exhaustion after 72 hours could explain the low concentration of MEL-related metabolites in the *SD method* (Figure 3-C). This was observed in *P. hubeiensis* (Konishi and Makino 2018), where an addition of olive oil was required every three days. We noted when FA were dosed every 24 hours, MEL production increased two fold when compared to the single dose regimen (Figure 3-) suggesting small doses of FA work better in this type of automated system. Moreover, it is possible that the presence of FA is required not only as building blocks of MELs but as an internal trigger for its production. An example of this was the observation for the SD regimen, in which, at the lowest dose and replenishing the media, the production of MEL-related metabolite was the lowest (Figure 3-). Nevertheless, this could also be attributed to a cell response, due to a constant stress forcing the cells from a nitrogen starvation state to a replenishing of nitrate, after each media supplementation.

The implementation of micro-fermentations for MEL production monitored over a 96 hours time course is the first of its kind, to our knowledge, for *P. graminicola*. We accomplished not only to standardise a detection method for MELs but also to quantify its production at low volumes.

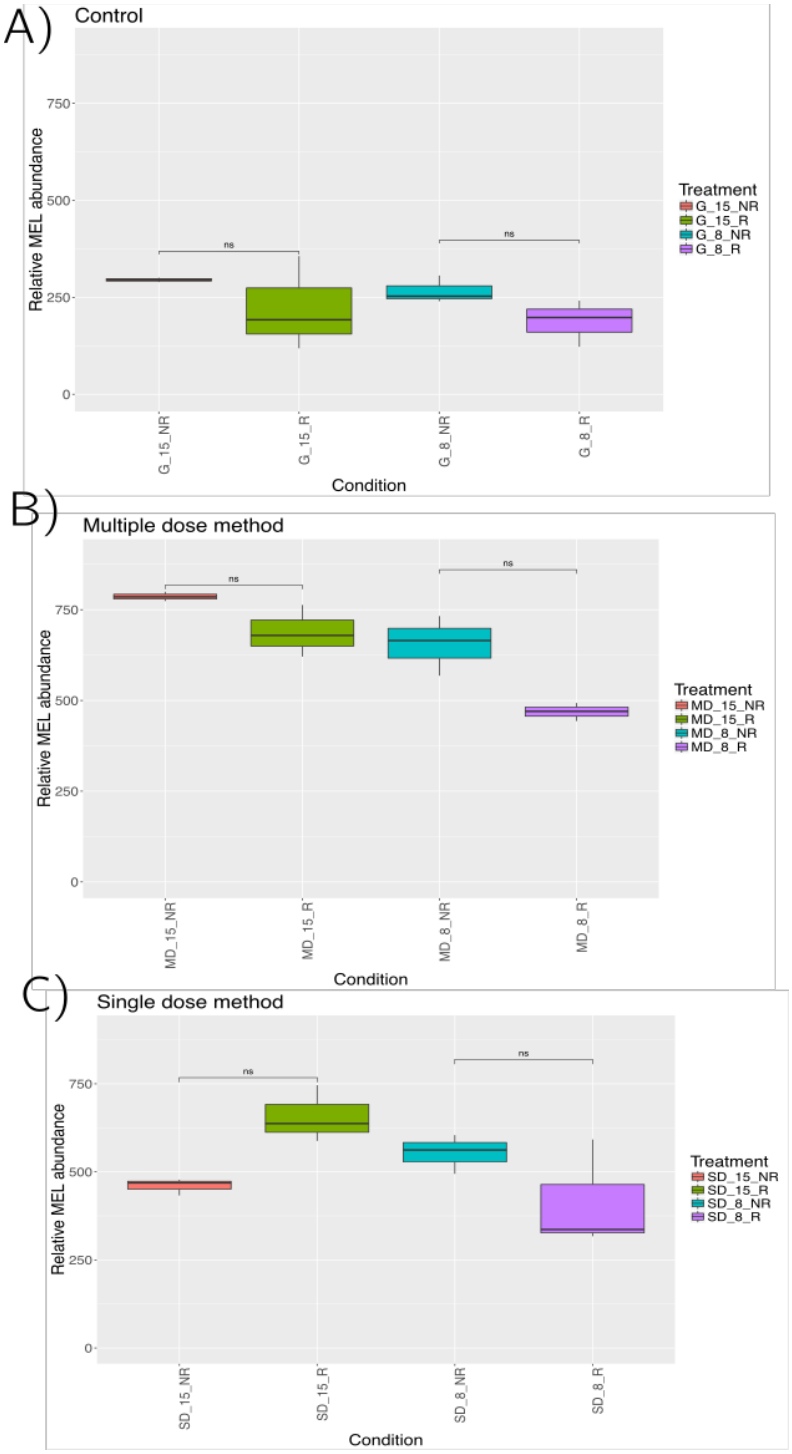



Figure 3-16. Relative abundance box plots for MEL related metabolites in micro-fermentations for *P. graminicola* after 72 hours. N=3. Showing treatments for A) control cultures with a final volume of 1500 µL (G_15) and 800 µL (G_8) with replace (R) and no-replace of media (NR). Same description for B) Multiple-dose (MS) and C) Single-dose (SD) regimes. Ns= non-statistically significant.

### 3.4.8 Comparison of MEL production using two different fatty acids as feedstock

The type of FA used as a feedstock affects the yield of MELs, even within the same strain (Kitamoto et al. 1990). Due to this complexity and variability in the production dynamics we aimed to compare different FAs sources for MEL production by *P. graminicola*.

#### 3.4.8.1 $^{1}$H NMR visualization from different sources of fatty acids

By analysing the sugar (5-3 ppm) and fatty acid region (3-0 ppm) of the NMR spectra, we observed that CRODAFAT and olive oil resulted in very different profiles when compared for the MEL enriched fractions. The olive oil has three sets of signals between 4-4.5 ppm (Figure 3-13,A), which are absent in the CRODAFAT. In addition, both FA lack a set of signals present in the MEL fractions in positions 4-3.5 ppm and 2.6-2.8 ppm (Figure 3-13,B,C)

As mentioned before, to discriminate between types of MEL (A-D) by $^{1}$H NMR is extremely difficult as their structures are very similar (e.g MEL-B and MEL-C having the same structure, being different only by the position of the acetyl group on carbon 4 or 6 respectively).

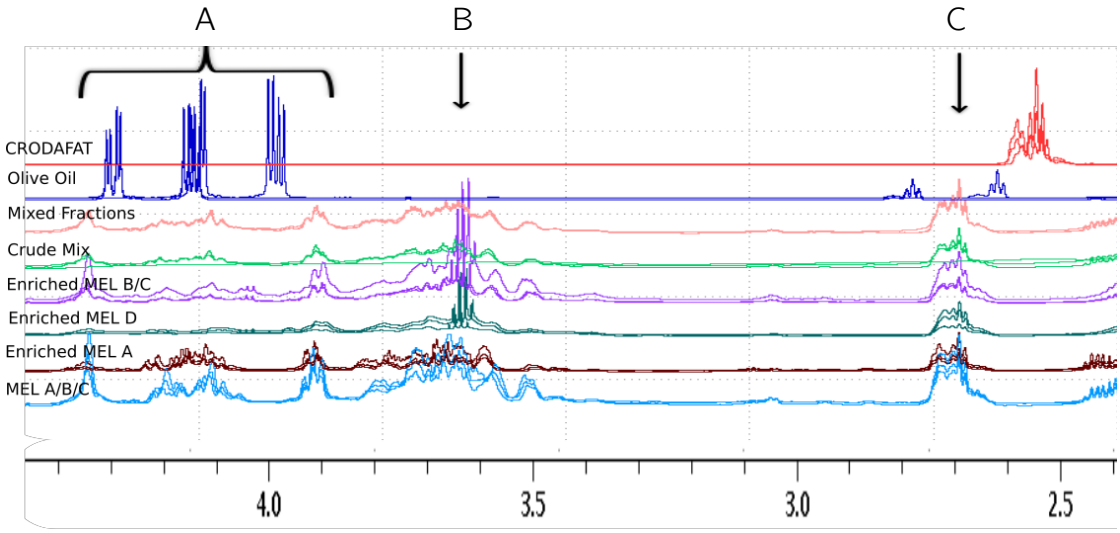



Figure 3-13. Mannose region of the $^{1}$H CPMG from different sources of fatty acid. NMR was conducted at 600 MHz. Data from three replicates is superimposed. Unique signals for olive oil at the 4.0 ppm region (A) absent in MEL fractions. Unique region regions associated with MELs (B and C), absent in CRODAFAT and olive oil.

#### 3.4.8.2 MEL production using CRODAFAT and olive oil as feedstock in P. graminicola wild type strain

In order to compare the MEL production with other studies, we replicated the feeding system implemented by Konishi and Makino (2018) where they fed olive oil to *P. hubeiensis* (another MEL-C producer) at the third and fifth day over seven days on a 5-L table-top fermenter. We tested olive oil and CRODAFAT as the FA sources by implementing a feeding system in which 0.11 grams of FA were added to the 1500 µL well from the beginning and the same dose was applied at the third day of fermentation. When comparing the MEL production between the CRODAFAT and olive oil feedstock, we observed higher production of MEL-related (Figure 3-18) and mannose-related metabolites (Appendix 3-8) with CRODAFAT. Interestingly unlike (Konishi and Makino 2018) who increased the MEL production in *P. hubeiensis* by adding a second dose of olive oil on the third day of the fermentation, we noticed a reduction, although an increase was observed when CRODAFAT was used. This could suggest MEL production is increased when saturated fatty acids are supplied to the media as CRODAFAT

has a high proportion of saturated fatty acids (Philippa Furnival, personal communication). The length of the carbon chain is similar for both, CRODAFAT and olive oil: 16-22 carbons.

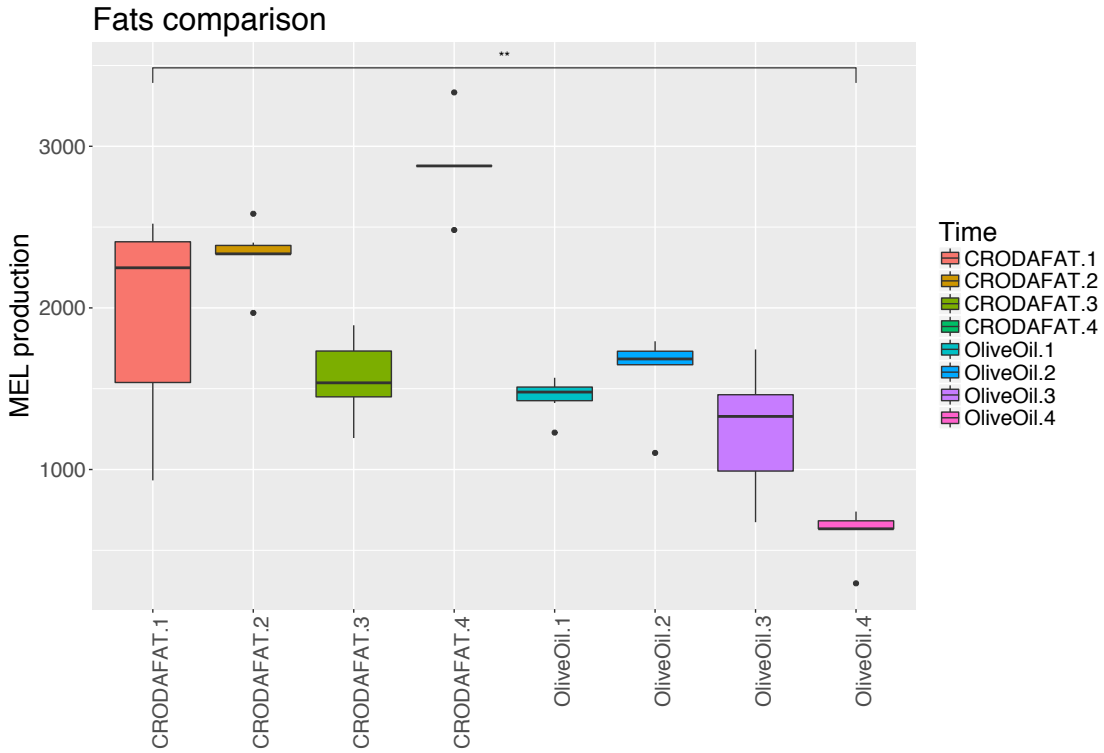



Figure 3-18. Comparison of MEL relative abundance from two different FA sources. Micro-fermentations for *P. graminicola* using CRODAFAT and olive oil as FA feedstock over a 96 hours time course. N= 6. Numbers on each condition refer to sampling time points: 1= 24 h, 2= 48 h, 3= 72 h and 4= 96 h. Displaying comparison between groups statistically significant, *: p <= 0.05 and **: p <= 0.01 or non-significant (ns).

## 3.5 CONCLUSIONS

We undertook and accomplished the development of a standardised protocol for production, detection and semi-quantification of MELs using *P. graminicola* CBS 10092 strain, without prior extraction and purification of this metabolite in a FA background, by $^{1}H$ NMR. We did this using CRODAFAT as the FA feedstock. This is as a mixture of waste FA residuals from industrial processes specific to CRODA, our industrial partner. To our knowledge this is the first report of MEL production for *P. graminicola* of this kind.

In terms of yield, as other published observations, we identified high variation in MEL production, dependent on both the fermentative conditions and FAs feedstock. We noticed in flask fermentation a single high concentrated dose of FAs yielded higher amounts of MELs than spreading that concentration in small multiple doses every 24 hours. On the other hand, for micro-fermentation systems, multiple small doses were more effective in terms of MEL yield than a concentrated single dose of FAs feedstock. The integration of evaluated variables (different well volumes: 800 and 1500 μL; open and closed system) in the micro-fermentation system raised the question if the constant stress for the cells passing from nitrogen starvation to nitrate refeeding might have cause the low yields. In future, a experiment using the same feeding system we proposed could be tested by adding media without nitrate, to evaluate the behaviour of the strain in terms of production.

In addition, the comparison of MEL production using two different FA feedstocks: CRODAFAT (mainly saturated) and olive oil (mainly unsaturated), resulted in a higher production with the former. It is possible that production of these metabolites might be favoured by the use of saturated fats over unsaturated fats. Consequently, variation in FA feedstock could shift the production towards different conformations of the MEL molecule (i.e carbon chain length, isomers).

Therefore, a standardised protocol that allows analytical quantification and characterisation of the MEL molecule is ideal. We attempted to use MS to characterise the MELs produced by *P. graminicola*. The complexity of the mixture, the lack of an optimised setup of the GeneMill metabolomics facility and the poor quality of standards, lead to low quality results. A priority for future work should focus on the optimisation of MEL isolation/purification from media, and MS characterisation in order to get an analytical quantification of its yield for the different forms of MEL produced.

We presented a comparison between different producing systems (batch and micro fermenter), however this has limitations. Is possible the physical properties of each system provoked inevitable variations on conditions such as aeration and pH, both being primarily important for MEL production dynamics.

In the next chapter we present the genomic annotation for *P. graminicola,* which serves as first step to understand the regulation and expression of the MEL cluster under induced conditions.

CHAPTER 4

# 4. THE MANNOSYLERITHRITOL LIPID (MEL) BIOSYNTHETIC CLUSTER IN *P. GRAMINICOLA* AND ITS REGULATION

## 4.1 INTRODUCTION

MELs are produced by most known members of the *Pseudozyma* genus. The first report of the MEL biosynthetic gene cluster was in the closely related fungus *U. maydis*, from which five genes were discovered to form the core cluster (Hewald et al. 2005).

### 4.1.1. MEL cluster proteins

The gene *emt1,* encodes a mannosyltransferase that produces mannosylerithritol, the sugar backbone of the MEL molecule, by mannosylation of erythritol (Figure 1-3). It was found to be essential for the production of MELs in *U. maydis;* its deletion incurred in a complete loss of the compound (Hewald et al. 2006).
*mac1* and *mac2* encode acyltransferases, which catalyse the transfer of fatty acids (short to medium chains) to the C-2 and C-3 positions of the mannosylerythritol (Figure 1-3). The biosynthetic pathway for MEL production in fungi is classed as a *chain-shortening* pathway (Figure 1-4A,B)(Kitamoto et al 1990). In mammals, an equivalent pathway is involved in the formation of bile acids, and similarly it takes place in the peroxisomes (Schultz, 1991). The elongation of the fatty acid (FA) chain is likely to occur via the β-oxidation pathway (Hewald et al. 2005, 2006; Teichmann et al. 2007) which also takes place in the peroxisomes. In many fungal species the synthesis of various secondary metabolites involves these organelles (Bartoszewska et al., 2011), such as production of the ß-lactam

antibiotic penicillin by *A. nidulans* and *P. chrysogenum*, which occurs in the peroxisomes (Müller et al., 1992; Spröte et al., 2009).

This is consistent with the identification of peroxisomal targeting signals (PTS) in both of these acetyltransferases from *U. maydis* (Freitag et al. 2014). Targeting either protein to the cytosol resulted in a lower production of MELs and a different carbon chain length, when compared to the wild type enzymes targeted to the peroxisome. These changes in the mis-targeted strains might be due to changes in the availability of the pool of acyl-CoA ester between organelles, being higher in peroxisomes (Freitag et al. 2014). This result allowed Freitag and collaborators (2014) to conclude not only the enzymatic function of MAC1 and MAC2 but also the intracellular location of the enzymes. Additionally, these enzymes cannot replace each other, as deletion of either *mac1* or *mac2* resulted in a complete loss of MEL biosynthesis (Hewald et al. 2006).

*mat1* encodes an acetyltransferase, catalysing the acetylation of the sugar moiety, which can occur at both positions C-4' and C-6' (MEL-A), at C-6' (MEL-B) or C-4' (MEL-C) alone or alternatively does not occur at all (MEL-D) (Figure 1-4B,C). In addition to this, their work showed this acetylation step is not essential for the secretion of the glycolipid, as *mat1* deficient strains secrete the deacetylated form of MEL: MEL- D (Figure 1-4B).

*mmf1* encodes a membrane transporter, which facilitates MEL secretion (Günther et al. 2015; Hewald et al. 2005; Konishi et al. 2010; Morita et al. 2006; Morita et al. 2007b). Mutants deficient for this protein were unable to produce extracellular MELs (Hewald et al. 2006). It was observed that the transporter cannot distinguish between MEL derivatives, as the spectrum of MELs, carrying acyl groups of different lengths, is quite broad (Hewald et al. 2005, 2006; Teichmann et al. 2007). This broad specificity is typical for members of the multidrug resistance major facilitators family (Del Sorbo, *et al.* 2000).

MMF1 displays high levels of sequence similarity to the gene *mFs1-1* from *Coprinus cinereus*, located in the region determining the mating type (Halsall et

al. 2000). This could indicate this transporter has a potential role in the function of the mating type locus, i.e for secretion of glycolipids which may enhance diffusion of hydrophobic lipopeptide pheromones (Hewald et al. 2005). Interestingly, with the exception of the transporter, homologues for the other four genes from the MEL cluster in *U. maydis* were found in *A. nidulans* (Hewald et al. 2006). These similarities between *U. maydis* and *A. nidulans* suggests a common evolutionary origin and possible horizontal gene transfer (Hewald et al. 2006).

Highly conserved orthologous MEL clusters have been identified in various *Pseudozyma* species including *P. antarctica* (Morita, Koike, et al. 2013; Saika 2014), *P. aphidis* (Günther et al. 2015; Lorenz et al. 2014), *P. tsukubaensis* (Saika et al. 2016), *P. hubeiensis* (Konishi et al. 2008; Sari et al. 2013) and *P. rugulosa* (Morita et al. 2006). In additional species its presence has been inferred based on the presence of MELs identified by chemical analysis (Deml et al. 1980; Kakugawa et al. 2002; Rodrigues; Konishi et al. 2007; Fukuoka et al. 2007, 2008; Morita et al. 2006, 2007).

### 4.1.2. MEL cluster: regulation

Expression of MEL cluster genes is induced by nitrogen starvation conditions (Hewald et al. 2005, 2006; Jezierska, Claus, and Van Bogaert 2018; Morita et al. 2014b; Nugent, Choffe, and Saville 2004). In addition, the carbohydrate source also plays an important role in the production of MELs. Based on relative transcript abundance, Morita et al (2014) observed lower expression of MELs cluster genes when *U. maydis* was cultured with soybean oil compared to when it was absent; whereas *P. antarctica* was able to produce the MELs under both conditions (presence and absence of FA). This variation on the regulation of MEL expression in response to FA in the media, indicates a dependency on the carbon source by *U. maydis*, being suppressed in oily conditions.

The source of FA also affects the yield of MELs. Use of coconut oil by *P. hubeiensis* KM-59 resulted in five times less MEL than if olive oil was used over a

four day fermentation (Konishi et al. 2008). Likewise, factors such as type of carbohydrate (glucose, mannose, erythritol, pentose among others) resulted in different yields of MEL (Athenaki et al. 2018; Konishi et al. 2008; Rau et al. 2005; Udo Rau et al. 2005; Sari et al. 2013). In these regards, there has been significant effort to understand how different FA sources affect the core MEL pathway function and expression (Fukuoka et al. 2008; Hewald et al. 2005, 2006; Morita et al. 2014b; Morita, Habe, et al. 2007; Morita, Koike, et al. 2013; Morita, Konishi, et al. 2007b, 2007a; Saika et al. 2016; Teichmann et al. 2007) but there is a need to identify and characterise the genes involved in the gene clusters regulation.

The preferred FA source for studies related to MEL production is soybean oil, however, some report the use of rapeseed, coconut, olive and sunflower oil (Fan et al. 2014; Isoda et al. 1997; Kitamoto et al 2002; Konishi et al. 2015; Medrzycka and Karpenko 2009; Morita et al. 2008, 2014b; Morita, Fukuoka, et al. 2013; Morita, Konishi, et al. 2007b; Rau et al. 2005; Yoshida et al. 2014). Very few investigations report the use of FA waste material as feedstock for MEL production, were waste frying oil or soap was used (Bednarski et al. 2004; Dzięgielewska et al. 2007). Our study CRODAFAT, an industrial biproduct, as the FA feedstock for MEL production.

## 4.2 Chapter aims

I. To produce RNA-seq data from fermentations under presence and absence of FA over a five day fermentation, in order to monitor MEL cluster
II. To produce qRT-PCR data from fermentations under presence and absence of FA over a five day fermentation, in order to monitor MEL cluster

### 4.2.1 Chapter description

In this chapter we describe the MEL biosynthetic cluster and utilise data generated from the genome sequence and annotation of *P. graminicola* (Chapter 3), which provided the identity and functional information of individual genes. There is not much information reported about the optimisation of growth conditions, therefore we used our industrial partner's growth and feed conditions. To the best of our knowledge there are no reports which extend over the full fermentation period. Therefore, from the integration of RNA-seq and qRT-PCR data we quantify the transcript expression of the MEL cluster genes during induced and non-induced conditions; using both fermenter and batch culture systems. This is the first report comparing these two producing systems for *P. graminicola* CBS 10092. This information helped us to identify that production in batch cultures is not very different from fermenter cultures. We also identified a potential preference of long carbon chains with respect to *mac1* transcription, regardless the system (batch or fermenter).

## 4.3 MATERIAL & METHODS

### 4.3.1 MEL cluster identification

In order to identify the MEL cluster genes, we used genomic and RNA-seq data for *P. graminicola*. We retrieved the protein sequences of the genes coding for the MEL cluster from eight related fungi (*Sporisorium reilianum* SRZ2, *Melanopsichium pennsulvanicum 4*, *Pseudozyma hubiensis* SY62, *Ustilago maydis* 521, *Ustilago hordei*, *Pseudozyma antarcticus* (recently renamed *Moesziomyces antarctica*), *Pseudozyma aphidis* DSM 70725 (recently renamed *Moesziomyces aphidis* DSM 70725), *Pseudozyma antarctica* T-34 (recently renamed *Moesziomyces antarcticus* T-34) and used this as database (For detail see Appendix. 4-). The sequences were retrieved from the NCBI website

(http://blast.ncbi.nlm.nih.gov/Blast.cgi). We used these as protein database against our gene prediction (Braker output[2]) and ran the *blastp* program from a terminal command line interface running the Linux version of x86_64 GNU.

### 4.3.2 *P. graminicola* MEL cluster phylogeny

Phylogenetic relationships were inferred using the amino acid alignments from ClustalW (Thompson *et al.* 1994) and processed with raxmlGUI software package version 1.5b1 (Stamatakis *et al.* 2008) by transforming the ClustalW alignment to a Phylip format. The analysis utilized the rapid bootstrap maximum-likelihood (ML), setting the bootstrap to 100 runs. The PROTGAMMALG model was used for the analysis of the alignment. The trees were drawn using FigTree and *A. nidulans* FGSC A4 as an outgroup.

### 4.3.3 *P. graminicola* ITS phylogeny

By using a ribosomal database (ITS including 5.8S rRNA) for multiple species, we looked for hits in *P. graminicola* genome. Once identified, we aligned all the copies found to check for variation and use the most representative sequence as input to construct a phylogenetic tree. The tree was constructed using a representative for each member from the aforementioned MEL cluster species database, by implementing ClustalW from command line, with 1000 bootstrap.

### 4.3.4 Culture conditions for P. graminicola

#### 4.3.4.1 Fermenter conditions (carried out at CRODA facilities)

Growth and sampling details were the same as section 2.3.3.1 and 2.3.3.2, respectively. Two independent fermentations were carried out, nevertheless, during

[2] See section 2.3.4

the second fermentation a failure on the air compressor from the bioreactor stopped the experiment after 72 h.

#### 4.3.4.2 Batch culture conditions

For *P. graminicola* batch cultures, a seed culture was obtained as 3.2.1. We used 1000 mL baffled flask containing 300 mL of producing media (Table 3-2) inoculated with 20 mL of the seed culture. For induced conditions we added 4 grams of CRODAFAT every 24 hours over the five day time course. Prior to adding the FA we took a 50 ml culture sample. Experiments were conducted in biological triplicates, each with three technical replicates.

### 4.3.5 RNA-seq: sequencing and transcript counts

The RNA extraction, sequencing library production and HiSeq platform sequencing was as described in section 2.3.3.

#### 4.3.5.1 RNA-seq data analysis

In order to get the total transcript counts for each library, we used the script HTseq-counts (Anders and Huber, 2016) that calculates the number of reads mapped for each transcript (Appendix. 4-). Read numbers mapping to each transcript were modelled, with negative binomial error distributions, using DESeq, which is an R package that identifies differentially expressed genes from raw count transcripts (Love et al 2014; Wang et al. 2009). Normalisation factors were calculated to correct for differences in library size among samples, which might otherwise cause bias in differential gene expression analysis by using the function *estimateSizeFactors*. This process makes the count values from each library comparable (Anders and Huber 2016). We implemented generalised linear models (GLMs) containing each of the three factors of interest (gene, time and condition) plus all two-way and three-way interactions. Common, trended and tag-wise

dispersion parameters were estimated. Tagwise dispersion was used for fold change estimating and significance testing. The estimated log2 fold change for each of the models and contrasts were tested in DESeq using a likelihood-ratios (LR) test (Wilks 1938). P- values associated with logFC (log2 fold change) were adjusted for multiple testing such that genes with a false discovery rate adjusted P-value < 5% were defined as significantly differentially expressed (Benjamini & Hochberg 1995). Pairwise comparisons of major interest (i.e. induced 24 h vs. non-induced 24 h; induced vs. non-induced) were also tested. To visualise whether and how overall patterns of gene expression separated samples by time, treatment or by replicates, a multidimensional scaling (MDS) plot was drawn using the plotMDS function in DESeq applied to all transcripts.

### 4.3.6 Quantitative real time PCR (qRT-PCR)

To monitor changes on the expression of the MEL gene cluster over time, under induced and non-induced conditions, we analysed cDNA samples from batch and fermenter cultures.

We DNAse treated 2 µg of total RNA, after this procedure, the RNA was mixed with 0.5 µL of hexamer primers, 0.25 µL of dNTP at 25mM and required water to reach a final volume of 5 µL. The mixture was incubated for 5 minutes at 65°C followed by 5 minutes incubation on ice. For the reverse transcription 0.5 µL of reverse transcriptase enzyme (Promega), 2 µL of 5X Buffer (Promega) and 1 µL of 100mM DTT were added to the mixture followed by 90 minutes incubation at 42 °C, 5 minutes at 72 °C and 5 minutes at 95 °C. The final product was visualised on agarose gel and quantified by nanodrop. Approximately 5-6 ng of the resulting cDNA was mixed with 2 µL of primer (1 µL of forward and 1 µL of reverse) at 8 µM, 10 µL of the SensiFAST™ SYBR® Hi-ROX Kit (Bioline) and water as required to reach a 20 µL volume. A 3-step cycling was performed using a StepOne® (Applied Biosystems) termocycler followed by 1 cycle of 2 minutes

incubation for polymerase activation at 95 °C and 40 cycles of 5 sec at 95 °C, 10 sec at 60 °C and 15 sec at 72 °C.

Gene expression levels were calculated relative to the expression levels of the constitutively expressed fungal gene encoding for actin over time (list of primers: Appendix. 4-). Relative expression was determined using the ΔΔCt method (Pfaffl 2001) and the values were expressed as percentage fold change. Noteworthy to mention, the batch samples used on this chapter were used in Chapter 3 to monitor MEL production by $^{1}$H NMR.

## 4.4 RESULTS & DISCUSSION

### 4.4.1 The *P. graminicola* MEL cluster

We identified the presence of the MEL cluster in *P. graminicola* from the gene calling and functional genomic annotation (2.4.6 and 2.4.7), confirmed it by amino acid homology and compared its genomic sequence to other MEL producers: *U. maydis* 521 and *P. aphidis DMS 70725*. From this comparison both the glycosyltransferase and the putative transporter share the highest similarities among the three species (Figure 4-1).

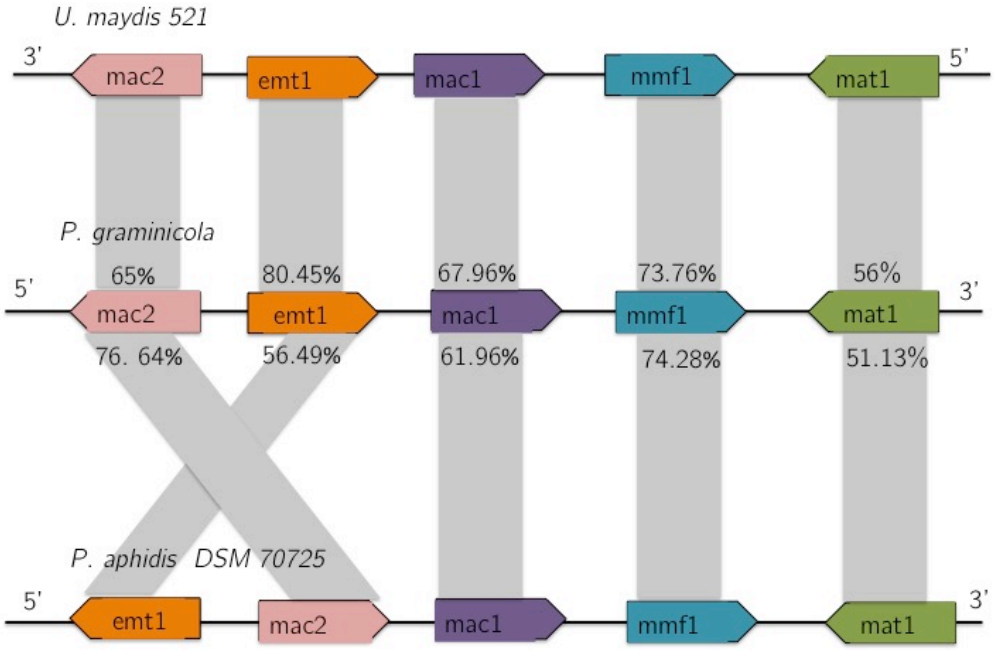



Figure 4-1. Alignment of the MEL biosynthetic gene cluster from *P. graminicola*, *U. maydis* and *P. aphidis*. The gene organisation of the three gene clusters are represented. Amino acid sequence comparison of *P. graminicola* MEL cluster genes to *U. maydis* 521 and *P. aphidis* DSM 70725 strains are given as the % identity as determined using the Blastp program for sequence similarity.

In order to evaluate the MEL cluster gene arrangement and location within the selected species, we took a region of 10 kb up comprising the MEL cluster boundaries. We identified the order and orientation of the MEL cluster genes is the same between both MEL-C producers (*P. graminicola* and *P. hubeiensis* SY62) despite the differences in the phylogenetic positioning. Nevertheless, *P. hubeiensis* SY62 has a gene between *emt1* and *mac1*, which is absent in *P. graminicola* and other *Pseudozyma* species. This gene is predicted to function as a choline phosphate cytidylyltransferase (Konishi et al. 2013). Curiously *U. maydis* also has an extra gene within the cluster in a similar location but this lacks any reported function; and both genes are orthologues in the respective species, sharing 50% identity, $e^{-0.95}$, 10% coverage of their amino acid sequence).

As a subsequent step we calculated the d-pairwise distances for the amino acid sequences between *P. graminicola* MEL cluster and our fungal database.

Consistently, *Sporisorium* genus showed the lowest values overall followed by *M. pennsylvanicum, Ustilago maydis* and *Pseudozyma* species (Table 4-1). The lower the values the higher the similarity between the sequences. The enzymes showing the highest and smallest variation among the comparison are MAT1 and EMT1, respectively. This suggests that the mechanism by which erythritol is mannosylated is more constrained and consequently the genes better conserved over the genus, compared to the acetylation process. This is perhaps not surprising, as the acetyltransferase can act at two different hydroxyl groups (C4 and C-6, Figure 1-4B), showing a relaxed regioselectivity (Hewald et al. 2006).

Table 4-1. Amino acid distances between *P. graminicola* MEL cluster sequence and other nine related basidiomycetes.

| SPECIES | EMT1 | MAC1 | MAC2 | MMF1 | MAT1 |
|---|---|---|---|---|---|
| *Sporisorium scitamineum* | 0.049 | 0.189 | 0.167 | 0.104 | 0.281 |
| *Sporisorium reilianum SRZ2* | 0.049 | 0.176 | 0.625 | 0.076 | 0.224 |
| *Melanopsichium pennsylvanicum 4* | 0.135 | 0.353 | 0.328 | 0.253 | 0.444 |
| *Ustilago maydis 521* | 0.14 | 0.307 | 0.36 | 0.209 | 0.386 |
| *Pseudozyma hubeiensis SY62* | 0.144 | 0.293 | 0.411 | 0.203 | 0.435 |
| *Ustilago hordei* | 0.155 | 0.371 | 0.299 | 0.253 | 0.421 |
| * *Pseudozyma antarctica* | 0.157 | 0.369 | 0.446 | 0.272 | 0.444 |
| * *Pseudozyma aphidis DSM 70725* | 0.161 | 0.363 | 0.442 | 0.274 | 0.446 |
| * *Pseudozyma antarctica T34* | 0.17 | 0.371 | 0.448 | 0.272 | 0.448 |

* genus recently changed to *Moesziomyces*

### 4.4.2 P. *graminicola* EMT1 protein

The glycosyltransferase gene, *emt1,* from *P. graminicola* encodes a protein of 617 amino acids and contains one intron of 90 nucleotides. The phylogeny of this gene shows significant divergence from other *Pseudozyma* species (Appendix4-4). By amino acid sequence comparison, this protein shows the highest level of identity to the *Sporisorium* genera and the phylogenetic positioning shows four main clusters: 1) the majority of *Pseudozyma* genera, 2) *Ustilago* species, 3) *P. graminicola* and *Sporisorium* plant pathogens together and 4) *Ustilago* maydis and

the MEL-C producer *P. hubiensis* (Appendix 4-4). In addition, by aligning the EMT1 amino acid sequences we identified a high level of identity for three regions interspersed with stretches of relatively low-identity, each of about 35 amino acids (Figure 4-2). We searched for sugar binding domain using the online tool CBS Pred, which is a carbohydrate binding site prediction (Malik et al. 2010) and the primary database Pfam (Finn et al. 2016) but this was not detected. The same three highly conserved regions were found in an alignment done by Saika and collaborators (2016).

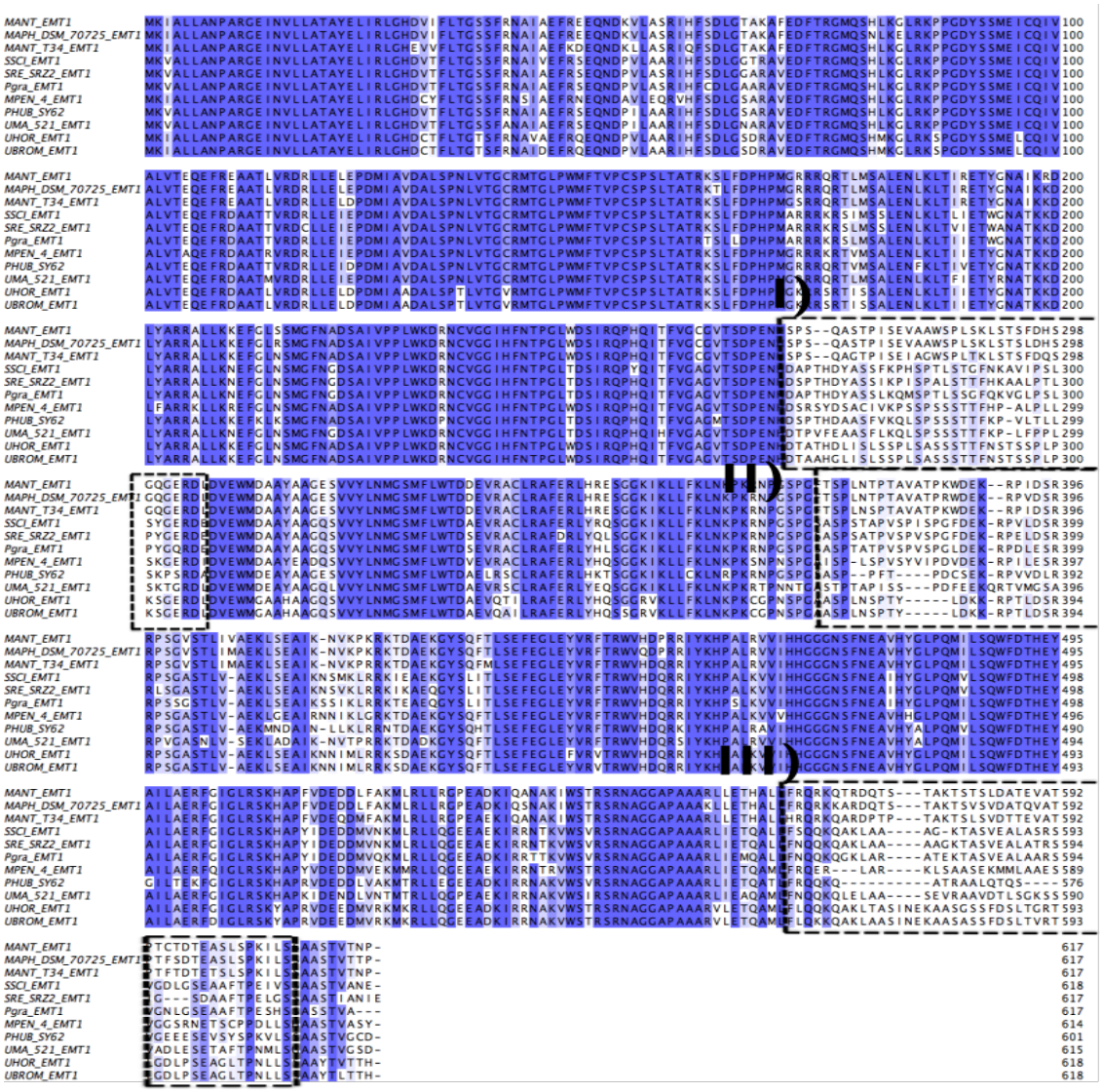



Figure 4-2. EMT1 amino acid alignment. Display of PgEMT1 and Basidiomycetes corresponding to the best blastp hits of PgEMT1. The alignment was done using ClustalW and for the display Jalview. Box coloured based on the percentage of similarity. Dotted boxes correspond to low-identity regions.

### 4.4.3 *P. graminicola* MAC1 & MAC2 proteins

The protein sequence for the acetyltransferase, MAC1, from *P. graminicola* (PgMAC1), consists of 585 amino acids and the gene includes no introns, whereas MAC2 consists (PgMAC2) of 551 amino acids and the gene has one intron of 103 nucleotides. The two enzymes share a low amino acid sequence similarity, despite fulfilling similar functions (21.70 % identity, $2e^{-12}$). In addition, their phylogenetic positioning locates *P. graminicola's* MAC1 in a subgroup with *S. scitamineum* within a clade containing *S. reilianum* (Appendix 4-5) whereas for MAC2 the clade

excludes *S. reilianum* (Appendix 4-6). In both cases *U. maydis* and *P. hubeiensis* group together.

A study carried out by Freitag and colleagues (2014) confirmed that both acyltransferases required for MEL production are targeted to the peroxisome in *U. maydis*. This localisation is defined by the peroxisomal targeting signal 1 (PTS1) prototype "SKL" peptide, that functions as a general motif for peroxisomal import (Gould et al., 1989; 1990). This signal peptide appears to be a conserved feature and is required for efficient assembly of MELs (Freitag *et al*. 2014). In *U. maydis*, the PTS1 motif is located at the C-terminus of both MAC1 (Ala-Arg-Leu) and MAC2 (Ala-Lys-Leu), and both are also found in *P. graminicola* (Figure 4-3).

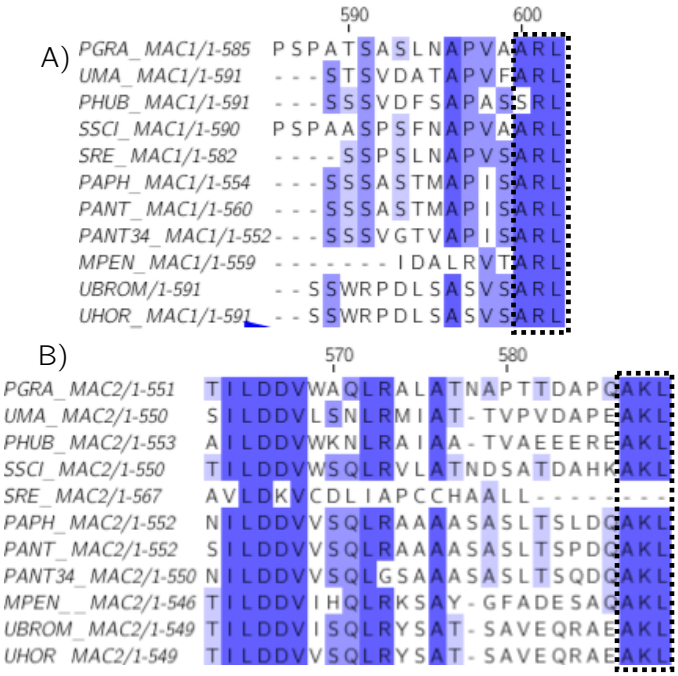



Figure 4-3. PTS1 signals for two acyltransferases. Display showing a section of an amino acid alignment for *P. graminicola* and Basidiomycetes fungi. A) Section of sequence alignment for best blastp hits for PgMAC1 showing **ARL** PTS1 motif. B) Section of sequence alignment for best blastp hits for PgMA2 showing **AKL** PTS1 motif. The alignments were done using ClustalW and for the display Jalview. Box coloured based on percentage of similarity. Dotted box corresponds to PTS1 signal.

### 4.4.4 *P. graminicola* MMF1 protein

The putative MEL transporter MMF1 for *P. graminicola* is the largest protein encoded within the MEL cluster, having 776 amino acids, and its gene has no introns. The corresponding phylogeny, when aligned to other basidiomycete orthologous, confirms the tendency of *P. graminicola* clustering to *Sporisorium* genera (Figure 4-4). For this protein we could not find a homolog in *A. nidulans* (Hewald et al. 2005, 2006), despite orthologues for the other genes from the MEL cluster have been found. However, this biosurfactant has not been characterised in *A. nidulans*. Interestingly *P. aphidis* (recently renamed *Moesziomyces aphidis)* behaved like an outgroup in the alignment.

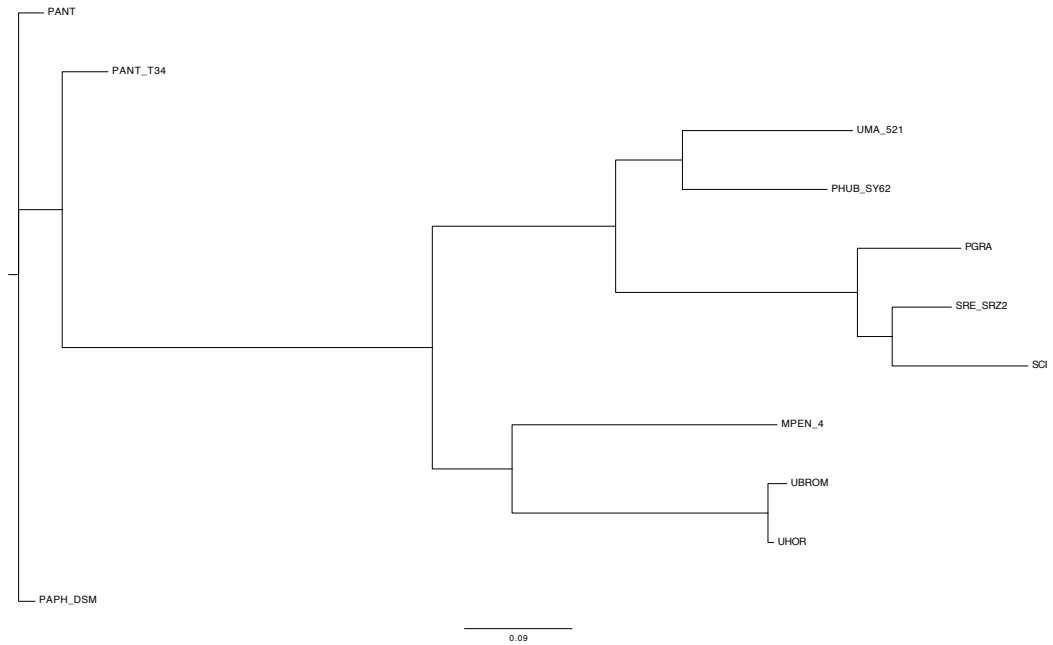



Figure 4-4. Phylogenetic analysis of MMF1. A molecular phylogenetic tree was constructed using the amino acid sequence of MMF1 for *P. graminicola* and other related fungi. Species key: UBROM: *Ustilago bominivora*, UHOR: *Ustilago hordei*, MPEN_4: *Melanopsichium pennsylvanicum*, 4, UMA_521: *Ustilago maydis* 521, PHUB_SY62: *Pseudozyma hubeiensis* SY62, PGRA: *Pseudozyma graminicola*, SRE_SRZ2: *Sporisorium reilianum* SRZ2, SSCI: *Sporisorium scitamineum*, PANT_T34: *Pseudozyma antarctica* T34, PANT: *Pseudozyma antarctica*, PAPH_DSM70725: *Pseudozyma aphidis* DSM 70725. The alignment was done using ClustalW. A rapid ML with a bootstrap of 100 runs was used. The three was drawn using FigTree.

This putative transporter is very well conserved in all the *basidiomycete* species included in this study (Figure 4-5).

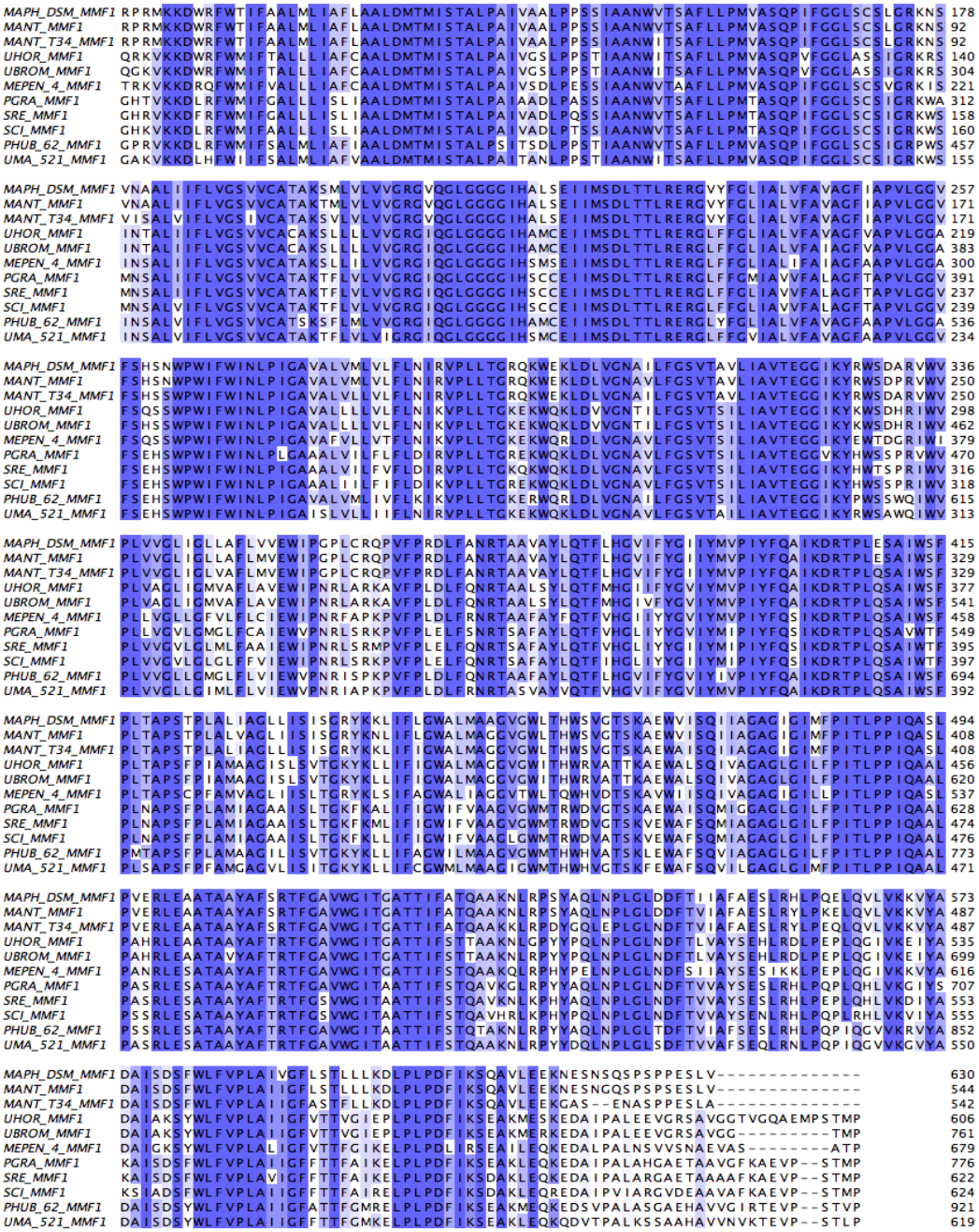



Figure 4-5. Amino acid sequence alignment of MMF1. Sequence displayed from Basidiomycetes corresponding to the best blastp hits of PgMMF1. The alignment was done using ClustalW and for the display Jalview. Box coloured based on percentage of similarity. High similarity observed overall the sequences.

### 4.4.5 *P. graminicola* MAT1 protein

The acetyltransferase for *P. graminicola* consisted of 531 amino acids and its gene contains three introns of 86, 95 and the biggest of 308 nucleotides. The phylogenetic positioning of this gene is very different when compared to the results from the other enzymes within the MEL cluster. Unlike the other four proteins, where *P. graminicola* grouped to the *Sporisorium* genera, MAT1 phylogenetic positioning did not show this (Figure 4-6). Instead, we observed this group split and *S. reilianum* was located at the root. Interestingly, *P. hubiensis*, which formerly clustered to *U. maydis*, forms an outgroup node acting as root for two sub groups; one comprising two *Ustilago* species and *M. pennsylvanicum* (MPEN) and other for the *Pseudozyma* species. *P. hubiensis* not clustering with *P. graminicola* was perhaps unexpected as both predominantly produce MEL-C, suggesting the acetylation pattern is not likely to work in the same way.

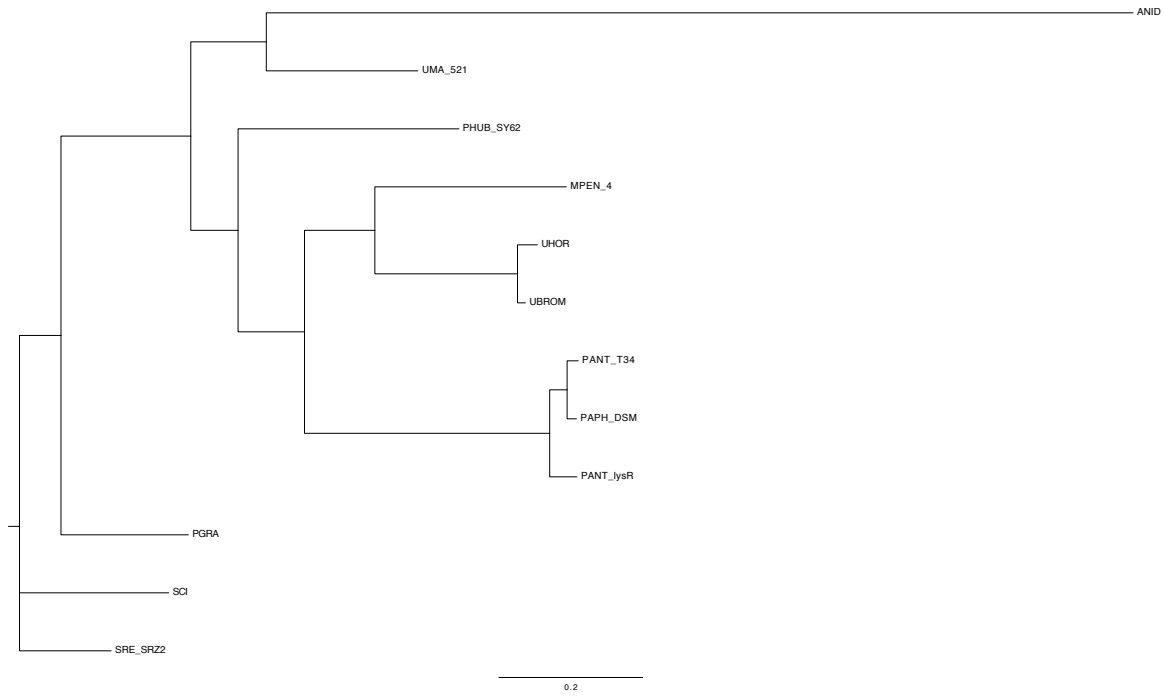



Figure 4-6. Phylogenetic analysis of MAT1. Amino acid sequence of MAT1 for *P. graminicola* and other related fungi. Species key: UBROM: *Ustilago bominivora*, UHOR: *Ustilago hordei*, MPEN_4: *Melanopsichium pennsylvanicum* 4, UMA_521: *Ustilago maydis* 521, PHUB_SY62: *Pseudozyma hubeiensis* SY62, PGRA: *Pseudozyma graminicola*, SRE_SRZ2: *Sporisorium reilianum* SRZ2, SSCI: *Sporisorium scitamineum*, PANT_T34: *Moesziomyces antarcticus* T34, PANT: *Pseudozyma antarctica*, PAPH_DSM70725: *Pseudozyma aphidis* DSM 70725, ANID: *Aspergillus nidulans* FGSC A4. The alignment was done using ClustalW. A rapid ML with a bootstrap of 100 runs was used. The tree was drawn using FigTree.

### 4.4.6 *P. graminicola* phylogeny: ITS and MEL cluster analysis

Conservation of the MEL cluster biosynthetic gene cluster across the *pseudozyma* genus, lead us to consider whether there was evidence for horizontal transfer. To address this, we first determined the phylogenetic relationship of the respective species. We utilised the proteomic sequences for *P. graminicola* and identified the orthologues to nine related smut fungi (Figure 4-7). We observed a high bootstrap for *P. hubeiensis* and *U. maydis* node. Additionally, *P. graminicola* clustered to *Sporisorium* species rather than other *Pseudozyma* (as confirm in Chapter 3). Nevertheless the bootstrap for this was relatively low (0.35), suggesting the resolution for that node was not very poor.

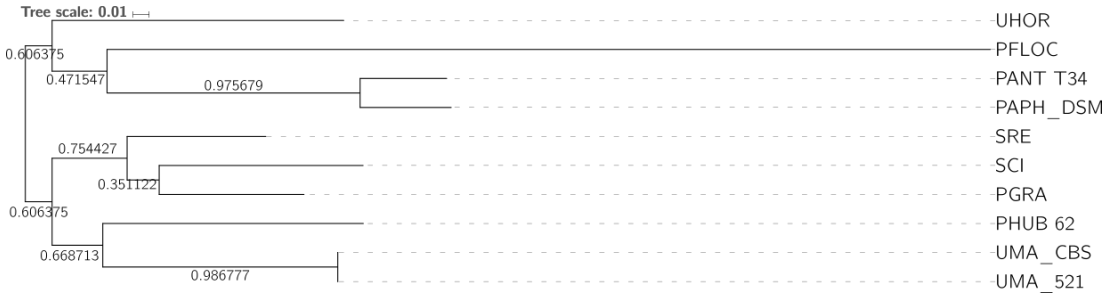



Figure 4-7. Phylogenetic analysis of orthologous proteins of *P. graminicola* to MEL producers. Bootstrap displayed on branches. Proteome sequences obtained from the NCBI. Species key: UHOR: *Ustilago hordei*, PFLOC: *Pseudozyma flocculosa*, PANT: *Pseudozyma antarctica T34*, PAPH: *Pseudozyma aphidis* DSM, SRE: *Sporisorium reilianum*, SCI: *Sporisorium scitamineum,* PGRA: Pseudozyma graminicola, PHUB: *Pseudozyma hubeiensis* UMA*: Ustilago maydis* strain CBS, UMA*: Ustilago maydis* strain 521. Analysis based on a total of 43815 genes.

For the MEL producers we created an equivalent phylogenetic specifically tree for the MEL cluster (Figure 4-8) that we compared to the orthologue proteins in order to infer potential gene transfer. From our results we observed a common ancestor for the MEL cluster, suggesting a low likelihood for an event of horizontal gene transfer taking place.

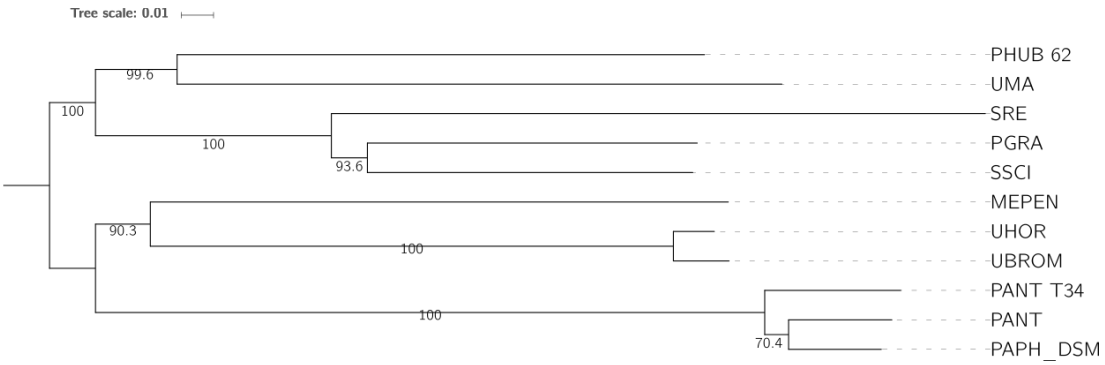



Figure 4-8. Phylogenetic analysis of MEL cluster amino acid sequences. Amino acid sequences of *Pseudozyma graminicola's* MEL cluster and best hits from blastp obtained from NCBI. Species key: PHUB_SY62: *Pseudozyma hubeiensis* SY62, UMA*: Ustilago maydis* 521, SRE_SRZ2: *Sporisorium reilianum* SRZ2, PGRA: *Pseudozyma graminicola,* SCI: *Sporisorium scitamineum*, MEPEN: *Melanopsichium pennsylvanicum* 4, UHOR: *Ustilago hordei*, UBROM: *Ustilago bominivora*, PANT T34: *Pseudozyma antarctica* T34, PANT: *Pseudozyma Antarctica,* PAPH_DSM: *Pseudozyma aphidis* DSM 70725. The alignment was done using Muscle with NJ applying JC model and 1000 bootstrap.

### 4.4.7 Putative Motif search: regulation of the MEL cluster?

We aimed to identify putative promoter motifs for the MEL cluster. In order to achieve this, we took a region of 1 kb upstream from the start codon of each gene from the cluster. Primarily, our approach involved the comparison of the equivalent regions from the *Basidiomycetes*: *U. maydis*, *P. antarctica*, *P. hubeiensis*, *S. reilianum* and the *Ascomycete A. nidulans* FSG 4 by aligning all the sequences. Nevertheless, this showed no obvious conserved sequence motifs.

Our second approach was to align orthologue genes from each one from the cluster in *P. graminicola* (i.e. *mac1* and *mac2* to their orthologues in the respective databased species*)*. By inspecting these alignments individually, we recognised the presence of GATA sequences, as reported by Hewald and collaborators (2005, 2006). However, the location of these showed no clear conservation across the species. Additionally, we implemented online tools such as MEME, which discovers patterns of sequences occurring repeatedly in a group of related sequences by displaying the probability of each possible nucleotide at each position (Bailey et al. 2009). From this analysis no clear motifs emerged.

Finally, we used the phylogeny results from our MEL cluster protein alignments, in order to identify the two closest related hits for each protein from *P. graminicola's* MEL cluster. We then performed pairwise alignments for each gene from the cluster and its corresponding hit, to identify potential conserved motifs. The motifs found by a pairwise alignment did not show any conservation when compare to the other alignments, suggesting the regulation mechanisms might not be well conserved.

Interestingly, analysis of the genomic region containing the MEL cluster we found a strong hit for a fungal transcription factor. The *P. graminicola* gene g6244 lies upstream from *mac2.* It encodes a putative protein with 53 % identity ($e^{0}$) to ASG-1, an activator of stress genes from *S. scitamineum* and 43 % identity ($5e^{-164}$) to the C6 transcriptor factor from *M. pennsylvanicum*. Interestingly, the orthologues for this gene in *U. maydis P. antarctica* and *P. aphidis* are not located at the boundaries of the MEL cluster, instead it is located 3 kb downstream. MEL expression has been linked to nitrogen starvation and presence of FA, this may imply that a stress response gene could be implicated in its regulation. In future we are aiming to delete this gene to check whether it is related to MEL regulation or not.

### 4.4.8 RNA seq data

We conducted twice, five day fermentations under both producing (with FA) and non-producing (without FA) conditions with samples being taken every 24 hours.

#### 4.4.8.1 RNA extraction yield

The extraction of RNA required of an optimisation process that involved two washing steps with PBS 1X, as the presence of FA affected the yield and the quality of the RNA. Our libraries showed a good size distribution and integrity, although the yield was unexpectedly low (Appendix 4-8) compared to the values

reported by the kit (approximately 450 µg, New Englad Biolabs). Despite this, we got enough material for the Illumina library preparation runs (Appendix 4-8 A,B and Appendix 4-9 A,B). The resulting cDNA libraries had an average length of between 300 to 550bp (Appendix 4-8C and Appendix 4-9C).

#### 4.4.8.2 RNA-seq data transcriptional variation

Our first approach was to identify patterns of gene expression separating samples by the presence or absence of FA or by the sampling time points. However, we did not observe a clear evidence of grouping between libraries, in relation to FA or time. In addition, we also identified a low correlation between replicates based on the distances separating libraries on the MDS plot (Figure 4-9).

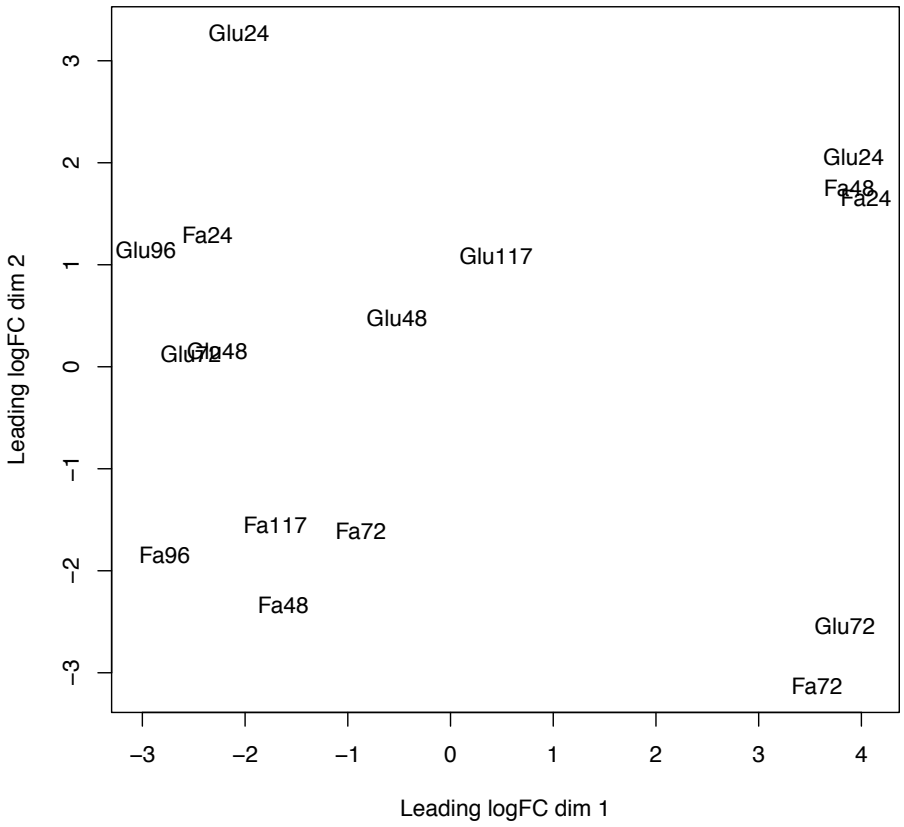



Figure 4-9. Multidimensional scaling plot (MDS) for *P. graminicola*. The log fold change in gene expression over a five-day fermentation under producing (Fa) and non-producing conditions (Glu) is presented. Number represents sample time point in hours (24,48,72, 96 and 117).

In order to measure the linear correlation between replicates, we constructed a binary pairwise comparison matrix between libraries and calculated the Pearson correlation coefficient ($r$) on the filtered counts (zeros removed). The coefficient values were used to plot a heatmap using R, from which a coloured bar ranging from 0.132 to 1 displays the level of similarity of expression profiles between samples. A perfect positive linear relationship would have an $r$ value of 1. The correlation levels did not follow a particular trend, regardless of condition or time point, even between replicates (Figure 4-10).

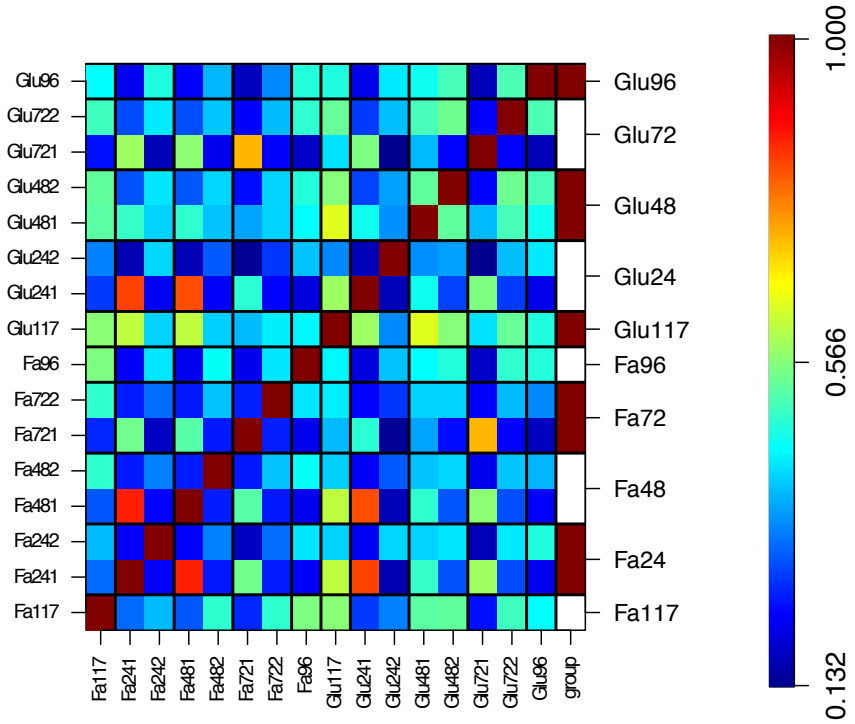



Figure 4-10. Pearson's correlation heatmap for *P. graminicola* gene expression. Samples grown under for producing (Fa) and non-producing (Glu) conditions. Number represents sample time point in hours (24,48,72, 96 and 117). The colour key (right) indicates the similarity of the sample to its replicate and ranges from low similarity (dark blue) to high similarity (red). Overall the transcriptomic profiles do not show a reciprocal correspondence between replicates.

#### 4.4.8.3 Statistical analysis: fold change and DGE

Secondary metabolite production are well known to have high intrinsic variation and dependency on external stimuli (Brakhage 2013). Nevertheless, the low

correlations observed between replicates might be mainly associated to the failure on the air compressor which potentially forced the cells into a stressful environment affecting the gene expression; rather than a merely intrinsic variation. Due to this, we focused our analysis towards a qualitative approach, rather than quantitative. Therefore, to identify key genes differentially expressed (DE) under the two conditions we filtered out genes that gave zero HTSeq-counts. We used the DESeq R package to normalise the filtered libraries counts using negative binomial distribution.

We observed the highest number of DE genes from the pairwise comparison between 96 hours to 24 hours, regardless the condition (Figure 4-12) and the majority of these DE genes corresponded to a down-regulation. Nonetheless, we only focused on the comparison between producing and non-producing conditions, which included all the libraries from each treatment. From this, we observed 63 DE genes, from which 44 were up-regulated and 19 down-regulated.

The up-regulated genes were related to functions for metabolic stress response, transport processes and G-protein receptor activated activity. These G-receptors are involved in activation of potassium and calcium channels, leading to a cascade of intracellular changes due to an altered cellular activity (Alberts et al. 2002).

The genes down-regulated coded for proteins belonging to the groups: hydrolases, ligases, transporters, nucleic acid binding, oxidoreductases, transferases, transmembrane receptors. Plenty of these proteins belong to the interleukin family, which regulates numerous biochemical events, such as cellular proliferation and long-term survival (Weaver et al. 2007). This data showed upregulation for only one of the expected metabolic groups for MEL producing conditions, transport processes, as observed on *P. aphidis* (Günther et al. 2015). However, the majority of up-regulated genes were involved in regulation and response of cellular activity. On the other hand, the genes down-regulated were involved in development and cellular growth.

We aimed to identify group of genes DE under MEL producing conditions associated to key metabolic functions, such as pathways involved in cell development, nitrogen and lipid metabolism (Günther et al. 2015).

Our differentially gene expression data did not show any clear trends for the transcription of the MEL cluster genes, nor did it show up-regulation of genes involved in lipid and/or nitrogen metabolism, as we would expect. One issue relates to irreproducibility, perhaps attributed to the faulty air compressor in the second fermentation, and this limits the biological interpretation of our results.

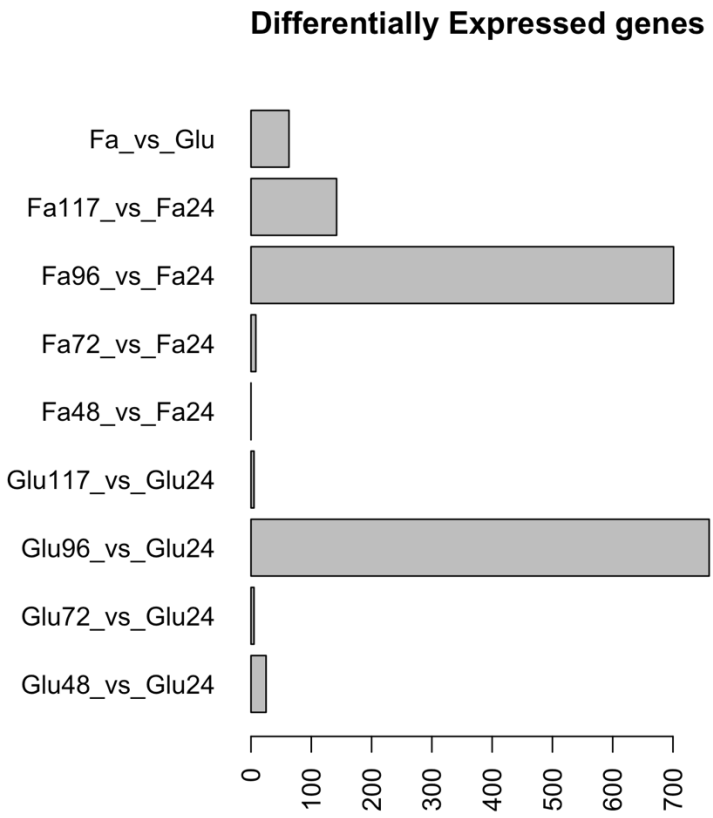



Figure 4-12. Differential gene expression in *P. graminicola* comparing fermenter cultures grown in the presence or absence of FA. Expression differences were inferred using generalized linear models (GLMs). The bars show numbers of genes that were DE between comparisons of time points within same condition and the upper bar shows the overall comparison between producing (Fa) and non-producing (Glu) conditions.

Then, we assessed the variation for gene expression between libraries by drawing scatterplots using the $\log_2$ from the transformed values of the normalised read

counts per gene. From these plots we expected to infer a relationship for gene expression between time points (from the same condition) but no apparent correlation was evident (Figure 4-13). Additionally, to identify a trend on transcriptomic expression, relative to producing conditions, we plotted the gene counts against the non-producing conditions. However, this comparison only showed few DE genes and did not demonstrate how the variable FA relate to changes in gene expression (Figure 4-13).

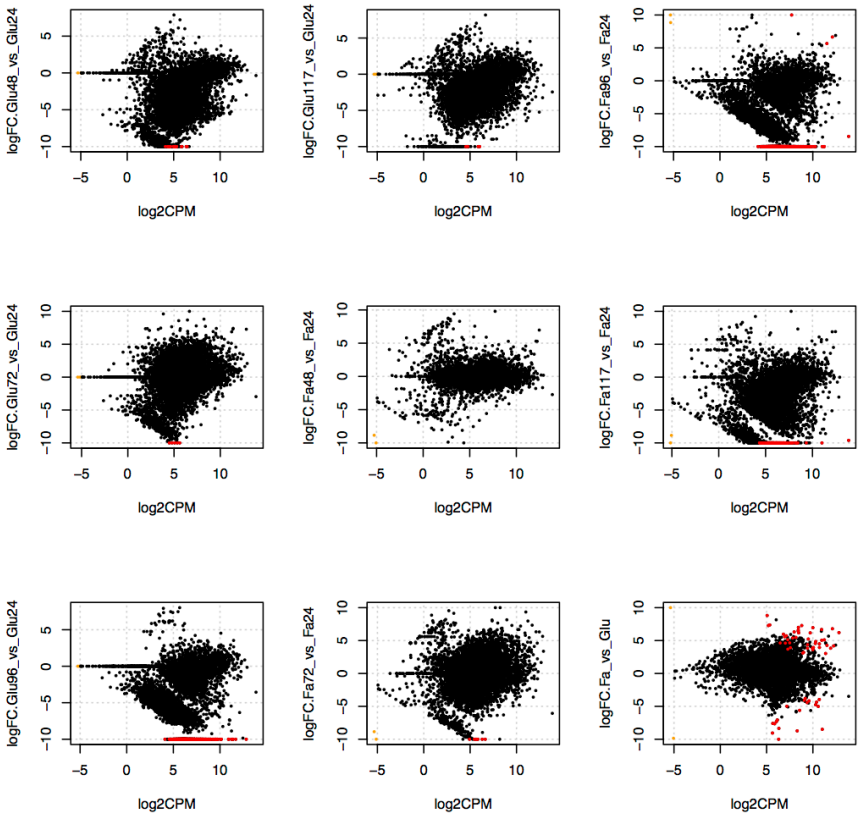



Figure 4-13. Pairwise scatterplot comparisons for transcriptomic expression for *P. graminicola*. Samples grown under producing (Fa) and non-producing (Glu). Significant DE genes with FDR-adjusted p-value < 0.05 highlighted in red. Black dots represent non DE genes. Transcripts tend to be down-regulated towards the end of the fermentation when time points are compared, regardless the condition. Nevertheless, when time points are pooled and the comparison is only between producing and non-producing conditions, some genes showed a clear differential expression.

The lack of DE genes was mainly attributed to a low number of replicates, which had an impact on the strength of the statistical analysis in addition to a low number of reads for some of the conditions (Chapter 2).

Nevertheless, in order to address this, we implemented qRT-PCR to monitor the expression of the MEL cluster genes, over the five day fermentation, using the same RNA samples as from the RNA-seq analysis. In addition to this, we emulated the fermenter conditions using batch cultures. The resulting data were used to calculate the fold change for each gene from the MEL cluster under both, producing and non-producing conditions for the fermenter and batch culture samples. We used the constitutively expressed gene actin to normalise the transcript levels.

### 4.4.9 qRT-PCR analysis for MEL cluster gene expression

#### 4.4.9.1 Generalised MEL cluster expression: batch and fermenter

In order to identify possible differences in gene expression common to both fermentative systems (batch and fermenter) over time, we combined the qRT-PCR data, to increase the number of replicates and thus the statistical power. The aim being to facilitate the identification of a characteristic pattern of gene expression during MEL producing conditions. Consequently, we plotted the mean of the fold change comparing with the following variables:

1) Condition: Control or treatment,

2) Source: Batch or Fermenter

3) Gene: *emt1*, *mac1*, *mac2*, *mat1*, *mmf1,*

4) Time-point: 24, 48, 72, 96 and 117 hours

Overall, we identified a high level of similarity in gene expression between both, non-producing and producing conditions (Figure 4-14A). We then, observed that the gene expression is slightly higher in the industrial system than in the batch

system (Figure 4-14B). Although this difference could be attributed to a higher human manipulation in the batch system than in the fermenter as the latter was automatised.

The highest fold change expression, regardless growth condition, was for *emt1*, followed by *mac1* (Figure 4-14C); whereas the other three genes from the cluster had a similar low expression. In *U. maydis,* where the MEL cluster has been well characterised, EMT1 catalyses the transfer of mannose on to erythritol, both sugars forming the backbone of the MEL molecule (Hewald et al. 2005; Teichmann et al. 2007). MAC1 and MAC2 both acylate mannosylerythritol with FAs of different lengths (Freitag et al. 2014). Additionally, 72 hours showed the highest mean for the MEL cluster gene expression (Figure 4-14D).

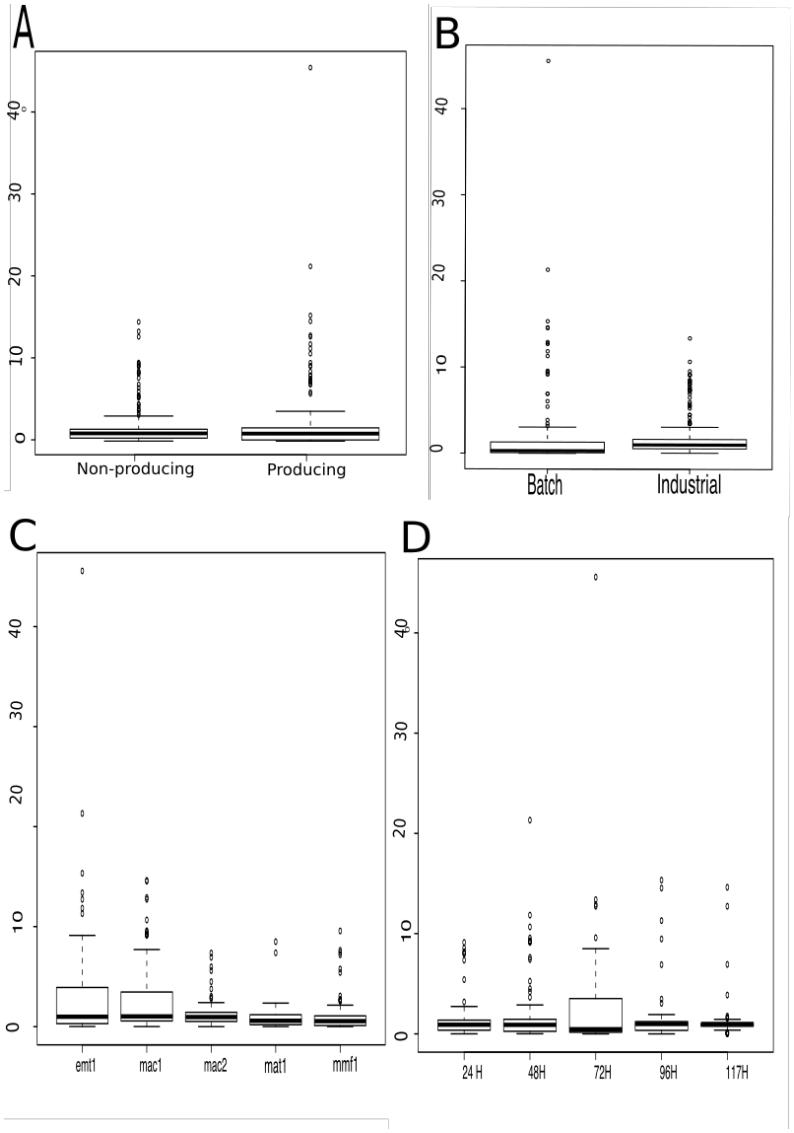



Figure 4-14. Mean boxplots for qRT-PCR data. A) Shows data distribution for non-producing (only glucose as carbon source) and producing (glucose and FA as carbon source). B) Data distribution for Batch and Industrial fermentation systems. C) Data distribution for MEL cluster genes regardless the condition. D) Data distribution for all MEL cluster genes over time regardless the condition. Box-plots display the mean, the first and third quartiles, and maximum and minimum values. Results from biological replicates on n = 3 (Batch system) or n= 3 (fermenter system) technical replicates.

We also aimed to assess fold change variation in response to a double combination of variables (Source and Gene, Time and Condition, Condition and Time). We observed that *emt1* and *mac1* had higher means for the fold change than the rest of the MEL cluster genes, which were expressed very similar over all the analysed combinations (Appendix 4-10). We identified 72 hours as the highest time point

for the MEL cluster gene expression, regardless the combination (Appendix 4-10B,C). This suggests a higher MEL production at this time point, being consistent with our $^{1}$H NMR semi-quantification of MEL production in batch system (3.4.6).

We aimed to identify the key variable or combination of them, which would explain the most the variation on the fold change for the MEL cluster. To achieve this, we implemented general linearized models (GLM), due to the multi-factorial nature of our experimental design. However, we could not identify a simple combination of variables (Appendix 4-11). Nevertheless, we can infer from the model that for some genes the time has a higher effect on the expression than the condition; implying the MEL cluster components are not co-regulated. Although, the necessity of such a complex model including interaction of more than three variables suggests either that we did not count with enough replicates or that we needed more variables to explain the variation in fold change.

### 4.4.10 MEL cluster gene expression: behaviour over time

We monitored the expression of the MEL cluster genes by qRT-PCR by implementing the ΔΔCt method (Livak and Schmittgen 2001). As the gene expression of the MEL cluster is induced by the presence of FA in *U. maydis*, *P. aphidis* and *P. antarctica* (Günther et al. 2015; Kitamoto et al. 1990; Morita et al. 2014a), we aimed to test this and to estimate the fold change for each transcript over a 117 hours time course under producing and non-producing conditions.

We calculated the percentage fold change for each gene over time, with respect to 24 hours and observed the gene *mac1* having the highest fold change values in comparison to the other four genes from the cluster (Figure 4-15A,B). The difference in *mac1* gene expression between producing and non-producing conditions, for the batch system, is approximately 10 fold. Interestingly in the absence of FAs, the genes *emt1*, *mac2*, *mat1* and *mmf1* seem to be co-regulated

(Figure 4-15A) with an abrupt overexpression of *mat1* after 48 hours.

On the other hand, in the presence of FAs the genes *emt1*, *mac2*, *mat1* and *mmf1* behave independently rather than as a co-regulated cluster (Figure 4-15B). *mac2* and *mat1* show a similar pattern of expression up to 72 hours; after this time the expression of *mat1* decreases dramatically, whereas *mac2* increases 5 fold. The glycosyltransferase *EMT1* has its highest peak of expression at 72 hours. After this time transcript levels decrease gradually over time (Figure 4-15B). Finally *mmf1* showed low levels of expression overtime, regardless the condition (Figure 4-15A,B). This is consistent with other investigations (Morita et al. 2008).

The expression of *mac1* gene under non-producing conditions was highest at 72 hours, and subsequently levels dropped dramatically (Figure 4-15A). Under producing conditions *mac1* gene expression increased gradually over time, reaching the highest expression values for any of the cluster genes (Figure 4-15B).

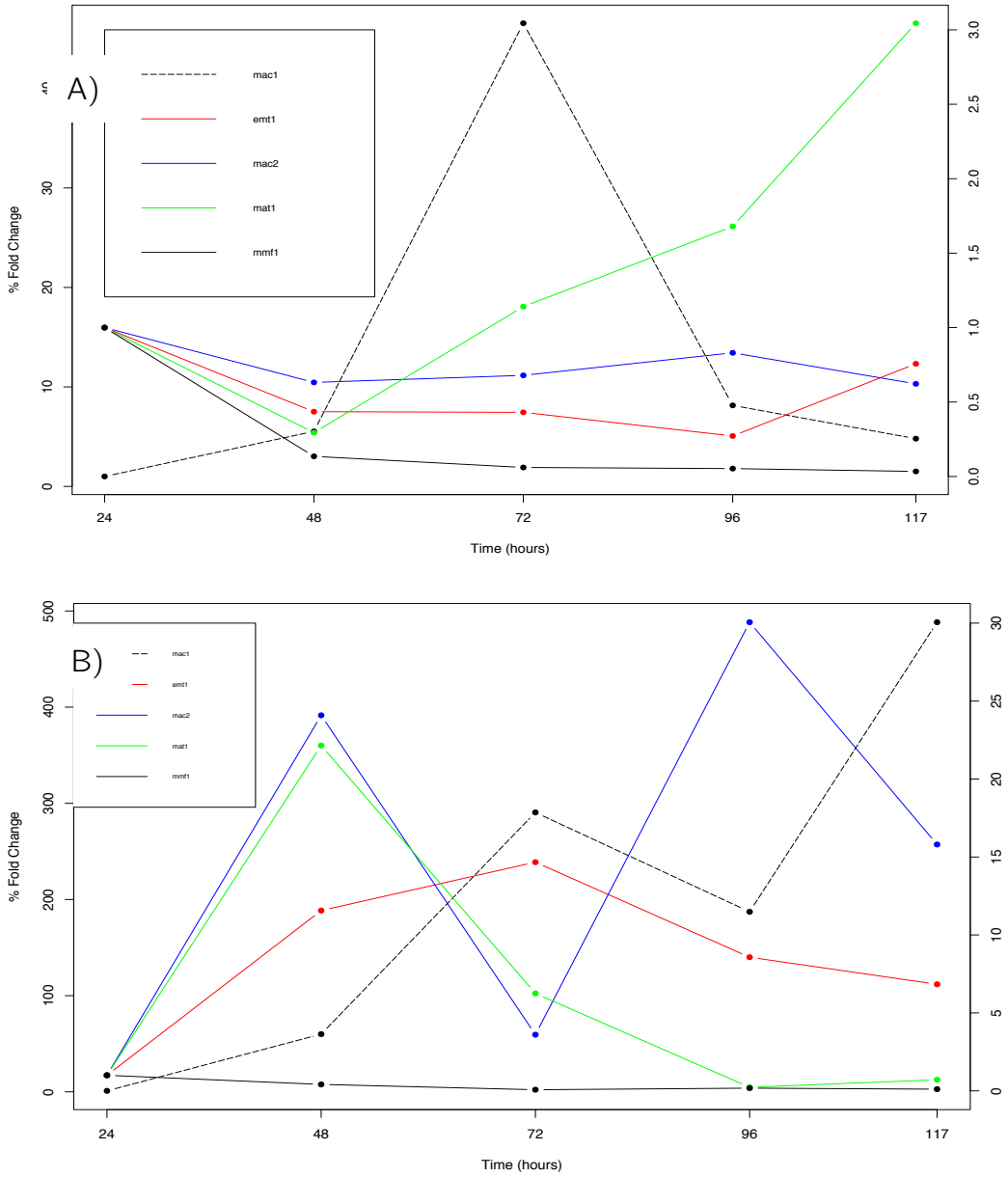



Figure 4-15. Fold expression for MEL cluster genes by *P. graminicola* batch cultures. Change over time for MEL cluster transcripts from batch cultures. qRT-PCR analysis was performed using the constitutively expressed actin gene (g6491) for normalisation. Relative expression was determined using the ΔΔCt method (Livak and Schmittgen 2001). Values shown are means of three biological replicates. A) Shows data for gene expression under non-producing conditions B) Shows data for gene expression under producing conditions. Left hand side scale for *mac1* gene(dotted line). Producing conditions: media supplemented with glucose and FA), non-producing conditions: media supplied only with glucose as carbon source.

In the fermenter, the transcription profile for the MEL cluster genes is completely different from that of the batch system; the genes seem to express as a co-

regulated cluster, whether FA were added or not to the media (Figure 4-16A,B). As with the batch cultures, *mac1* showed the highest values for gene expression in both producing and non-producing conditions.

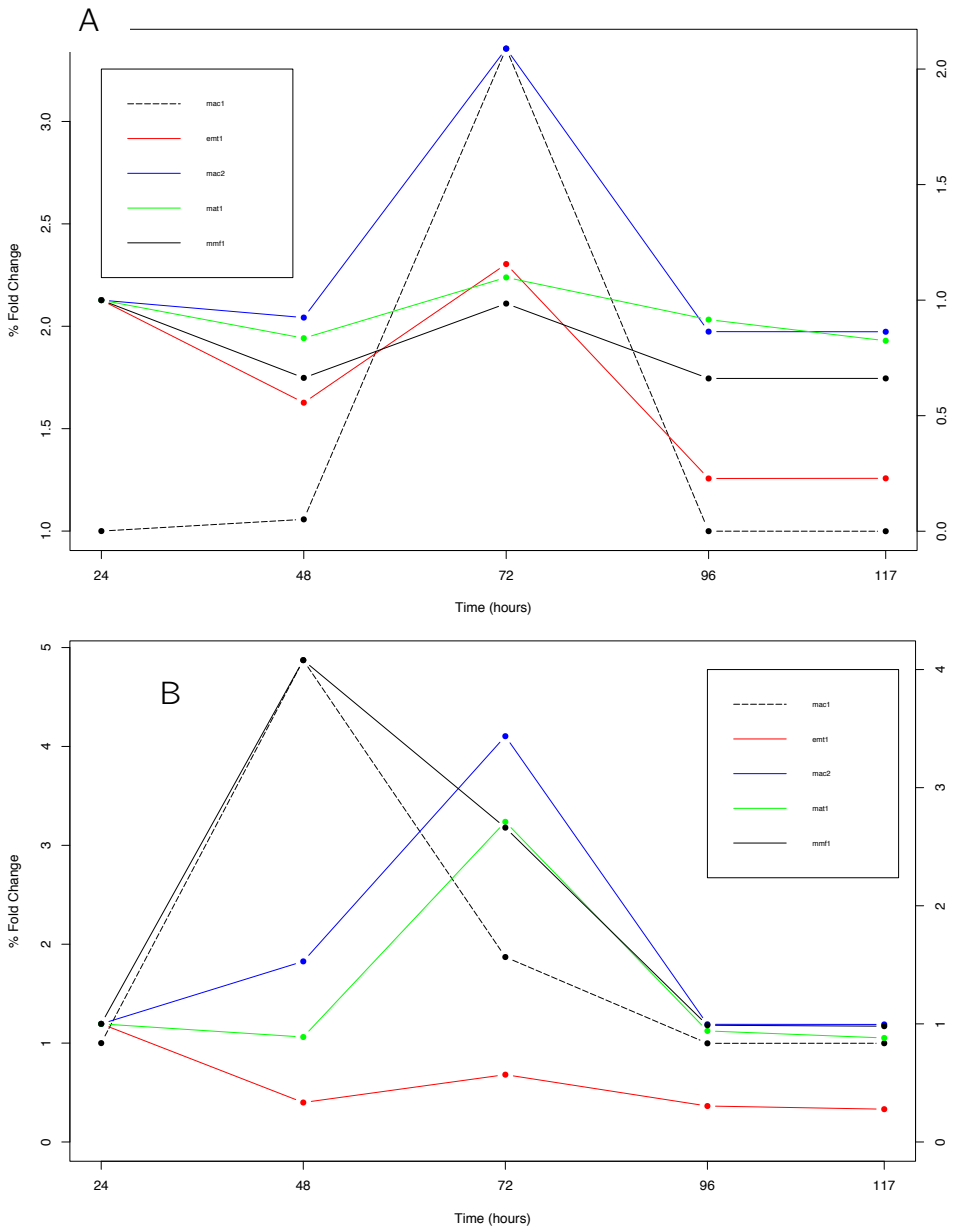



Figure 4-16. Fold expression for MEL cluster genes by *P. graminicola* fermenter. qRT-PCR analysis was performed using the constitutively expressed actin gene (g6491) for normalization. Relative expression was determined using the ΔΔCt method (Livak and Schmittgen 2001). Values shown are means of biological replicates. A) Shows data from non-producing conditions (only glucose as carbon source). B) Shows data for producing conditions (media supplemented with glucose and FA as carbon source). Left hand side scale for *mac1* gene (dotted line).

The expression of the MEL cluster genes is not induced by the presence of FA.

The non-reproducible data from our RNA-seq analysis was proven later by the qRT-PCR analysis from the fermenter samples, were the MEL cluster did not show a remarkable induction.

Regardless the fermentative system, we identified a lack of synchrony in the expression of the genes from the cluster, which strongly suggest they do not necessarily express as a block. Interestingly, batch cultures data for the transporter *mmf1* has the lowest expression values, suggesting a potential cause for the low yields reported for *P. graminicola* species. Additionally, we observed under the presence of FA, *mac1* is highly express, which might be attributable to a preference for long carbon chains (longer than 18 carbons) over shorter chains, as the CRODAFAT, used to feed the organism with, predominantly has palmitic ($C_{22}$), stearic ($C_{18}$), oleic ($C_{18}$), linoleic ($C_{18}$) and erucic acid ($C_{22}$), all having at least 18 carbons on their chain.

Freitag and collaborators (2014) tried to elucidate the position specificity for both acyltransferases with no success; although they found that both enzymes (MAC1 and MAC2) are targeted to peroxisomes and a mis-targeting to the cytosol arise in an altered acetylation pattern. They probed that this alteration has to due only with the localisation of the enzymes and not due to changes in the gene expression. This might support our supposition of *mac1* preference on long carbon chains over short ones as both enzymes are expected to work together (Hewald et al. 2005).

## 4.5 CONCLUSIONS

We demonstrated the presence of the MEL cluster genes in *P. graminicola* and observed that these are closely related to those of *Sporisorium* rather than other *Pseudozyma* species. We also observed, despite *P. hubeiensis* is also a MEL-C producer, its genome arrangement is different from *P. graminicola*, and it lacks the potential regulator, present in the *Sporisorium* and *P. graminicola* species. It is possible the regulation of the MEL cluster takes place in a different way than those in *U. maydis*, therefore comparisons to this species do not allow to unravel this type of information.

Alongside, our aim in this chapter was to identify DE genes under MEL producing conditions, focusing primarily on the MEL cluster; and secondly on genes related to key metabolic pathways. For this, we processed fermenter samples and analysed these by RNA-seq analysis. Using these data we determined that the presence of FA does lead to a general increase on the expression of the MEL cluster in *P. graminicola*, although a clear coordinated response was not observed.

To investigate this further we utilised qRT-PCR analysis, to monitor the expression of the MEL cluster during producing and non-producing conditions. We also conducted a qRT PCR experiment using samples from batch cultures.

Unlike *U. maydis* and *P. aphidis*, the MEL cluster genes in *P. graminicola* do not appear to be co-regulated. This might indicate that the MEL cluster genes are involved in roles other than the production of these secondary metabolites (Hewald et al. 2006; Tollot et al. 2016). In addition, despite the limitations from the RNA-seq data in our study, the integration of *omic* techniques allowed us to determine the behaviour of the MEL gene cluster do not necessarily meet the production of the compound. From our $^{1}$H NMR we observed that 72 hours in some of the fermentative systems yielded the lowest concentration of MEL, whereas our qRT-PCR analysis shows at this time point the highest values for gene expression, under producing conditions.

Even when MELs gene cluster expression has become a targeted topic to investigate, the vast majority of studies related to its production come from shake flask cultivation; only few reports exist using bioreactor production (Adamczak and Bednarski 2000; Kim et al. 1999; U. Rau et al. 2005) and according to the extend of our knowledge, so far, this is the first study which reports 1) a comparison between both systems, batch and bioreactor and 2) a time-course analysis of gene expression. Most of the reports (up to the writing of this document) use an end-point approach, were a comparison between the start and end of the fermentation takes place (Morita et al. 2014b; Morita, Konishi, et al. 2007b; Saika et al. 2016; Yoshida et al. 2014).

On the other hand, from literature review, when a time course is reported instead of a genetic expression analysis, an analytically directed one takes place (Faria et al. 2014; Günther et al. 2015; U. Rau et al. 2005; Udo Rau et al. 2005). This makes our results novel and useful to undertake future complementary experiments as sets the ground for optimisation, limitations and scope.

As an integrative analysis to validate our annotation and to expand our understanding of MEL production, in the next chapter we undertook the creation of deletion mutants for genes potentially involved in MEL biosynthesis and regulation system. We monitored the affect these mutations had on the production yield and identified the morphological changes these mutants had, to gain insights into their biological roles.

# 5. ANALYSIS OF PUTATIVE REGULATORS FOR MEL PRODUCTION

## 5.1 INTRODUCTION

The availability of genomic data for MEL producers has allowed studies targeting the expression of the MEL biosynthetic cluster and the manipulation of genes, contributing to the understanding of MEL synthesis (Flagfeldt et al. 2009a; Günther et al. 2015; Konishi et al. 2015; Konishi and Makino 2017; Saika et al. 2016). A promoter is defined as a region upstream the gene to be transcribed where the RNA polymerase complex attaches, in order to initiate transcription (Mudge and Harrow 2016). Therefore, the identification of promoters associated to regulation of the MEL cluster corresponds to a good approach as first step to optimise the production.

Hewald and collaborators (2006) were the first to compare the promoter regions for the MEL cluster in *U. maydis* and they found no obvious conserved sequences between this fungus and other *ascomycetes*. However, they identified several GATA sequences in the putative promoter regions for the MEL cluster gene, *emt1*. This finding suggest a potential role for the GATA factor homologue to the general nitrogen regulator *areA* from *A. nidulans* (Hewald et al. 2005) in the regulation of this gene cluster (Hewald et al. 2006; Macios et al. 2012).

Günther and collaborators (2015), using RNA-seq analysis of *P. aphidis* under MEL producing conditions (presence of soybean oil), identified five main groups of genes up-regulated; among these, a group of transcription factors and regulatory proteins, were particularly interesting. The first, a transcriptor factor (PaG_100136) had a Myb-like DNA binding domain and is likely to be involved in regulating growth and development (Ness, 1999). The second, a zinc finger

transcription factor (PaG_00705), which may be involved in regulation of glycolipid synthesis (Günther et al. 2015). Finally, the gene PaG_05062, which had strong homology to the *U. maydis* regulatory protein PAC2 and is known to inhibit hyphal growth (Elías-Villalobos, Fernández-Álvarez, and Ibeas 2011).

Interestingly, Tollot *et al*. (2016) investigating the function of a regulator of sporogenesis (*ros1*) in the pathogenic development of the fungus *U. maydis*, found this gene up-regulates the MEL cluster. This gene is essential for teliospore production at the late stage of the biotrophic life cycle of *U. maydis* (Tollot et al. 2016) and sporulation is commonly associated to secondary metabolism (Calvo *et al.* 2002). The *ros1* gene is a member of the WOPR family, a novel class of fungal specific transcriptional regulators, that binds DNA via their N-terminal WOPR box, which in most of fungal genomes has two paralogues (Kunitomo et al. 1995; Caspari 1997). This WOPR family has been extensively studied in *ascomycetes*, having a conserved function in the control of developmental processes (Tollot et al. 2016) such as plant colonization and sexual/asexual reproduction in phytopathogenic fungi (Michielse *et al*. 2009a, 2011b; Jonkers *et al*. 2012; Santhanam *et al*. 2013; Mirzadi *et al*. 2014; Okmen *et al*. 2014). Another member from the WOPR family is Wor1, the master regulator of the white-opaque phenotypic switching allowing *Candida albicans* to adapt to niches in the human host, therefore is very well characterise (Lohse et al. 2006).

## 5.2 Chapter aims

I. To construct deletion mutants for potential genes involved in regulation of the MEL cluster in order to identify key genes that help to improve MEL production

II. To monitor the MEL production of the deficient mutants by $^{1}$H NMR

### 5.2.1 Chapter description

In this chapter we described the construction of deletion mutants for *P. graminicola* wild type (WT) strain. We report the impact of *emt1* gene deletion on the production of MELs, monitored by $^{1}$H NMR. In order to construct this mutant, we developed a successful transformation protocol using *U. maydis* selection cassettes and promoters. By homologous recombination we knocked out three other genes with a putative regulation function for MEL production: *areA,* involved in the regulatory response to nitrogen availability. *pac2,* the orthologue of which functions as a potential MEL regulator in *P. aphidis* (gene PaG_05062) and the paralogue, *gti1*, both of which are homologous to the *U. maydis ros1.*

## 5.3 MATERIALS & METHODS

### 5.3.1 Construction of deletion mutants for *P. graminicola* CBS10092 strain

#### 5.3.1.1 Plasmids and knock out plasmid construction

We used the plasmid pMF1-c (Appendix 5-1) carrying a carboxin resistance marker and a hygromycin resistance marker, genes Cbx$^{r}$ and Hyg$^{r}$, respectively, both driven by the arabinase gene promoter ($P_{crg1}$) for *U. maydis* (Brachmann et al. 2004). The construction of deletion cassettes was based on Gibson cloning (Gibson et al. 2010) method and NEbuilder (NEB) reaction mix. For this, we use primers (Appendix 5-) to amplify the flanking regions 1 kb upstream and downstream from the open reading frame (ORF) of the targeted genes (Figure 5-1A) and for the resistance genes from the plasmids. The primers used introduced complementarity to the tails (5' and 3' ends) of the respective PCR products (Figure 5-1B), which were joined with NEB builder mix by this complementarity (Figure 5-1C) to form the deletion cassettes. These cassettes, were cloned into pMINIT and introduced into *E. coli* using the PCR cloning kit (NEB), following manufacterer instruction's

(Figure 5-1D). Transformants carrying the plasmid were selected on LB ampicillin plates. DNA was extracted from transformants using the GeneJet Plasmid miniprep (ThermoScientific) following manufacturer instructions (Figure 5-1E). High-fidelity polymerase Phusion® (NEB) was used to amplify the deletion cassette (FRs joined to resistance marker). The PCRs were run on agarose gel and eluted using the QIAquick PCR purification kit (Qiagen). The final product was quantified by nanodrop (Figure 5-1F) and used to directly transform *P. graminicola* WT (Figure 5-1G).

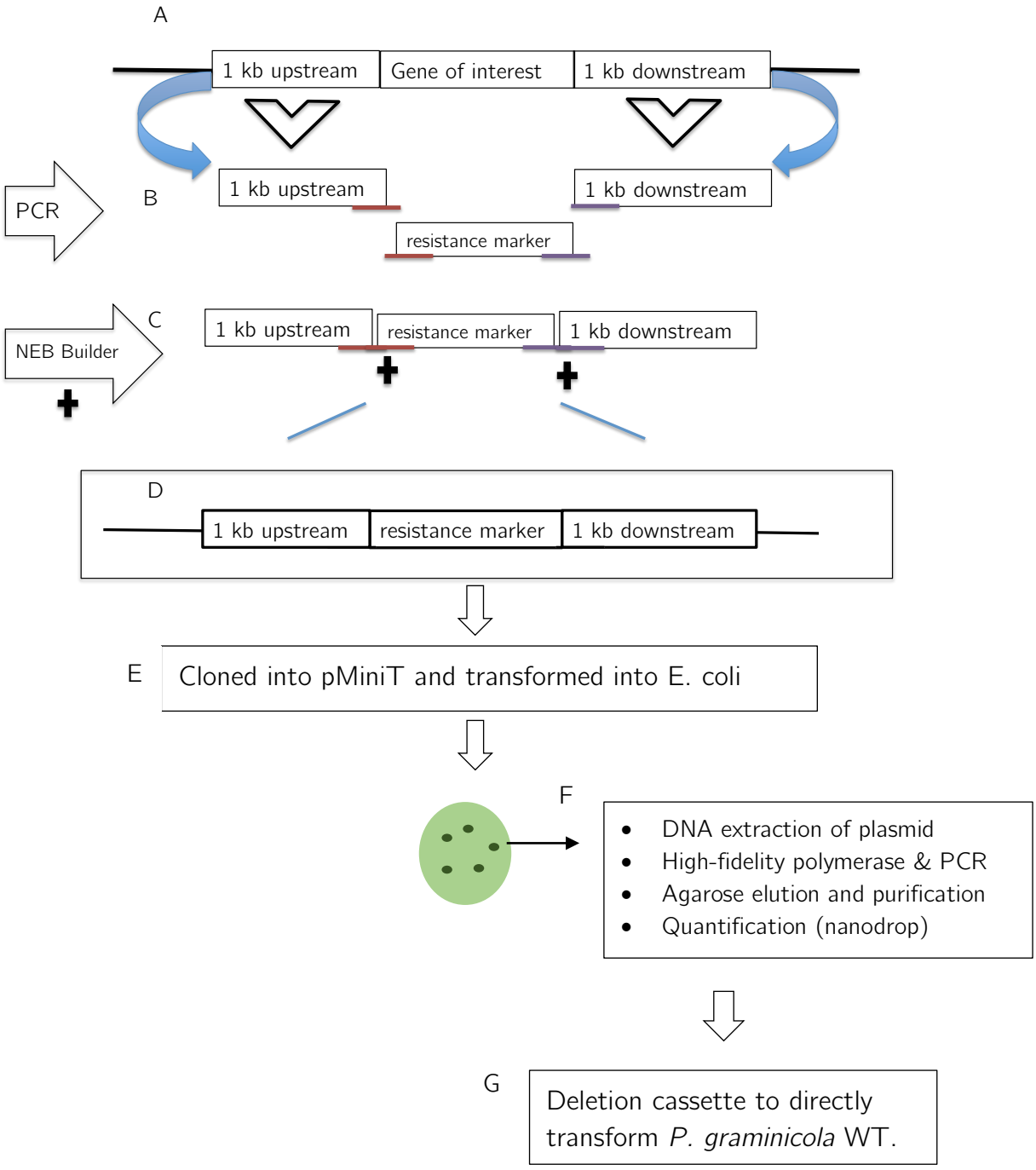



Figure 5-1. Diagram for the construction of the deletion plasmids used to transform *P. graminicola* WT strain. A) Representation of genomic region from P. *graminicola* WT to amplify flanking regions (FR) by PCR, is depicted. B) The PCR fragments amplified carry complementary nucleotides between the FR and the resistance marker gene (red and purple 5' and 3' region respectively). C) Using the NEB builder kit ® the FRs, amplified from *P. graminicola,* and the resistance marker gene were joined. D) Deletion cassette formed which E) subsequently was cloned into pMiniT and trasformed into DH5 alpha *E. coli* competent cells. F) DNA was extracted from plasmids and Phusion polimerase was used to PCR-up deletion cassette. The PCR product was run on agarose gel, eluted and purified using a commercial kit. G) The final product was quantified by nanodrop and directly used to transformed *P. graminicola* WT strains.

### 5.3.2 *P. graminicola* wild type transformation procedure

#### 5.3.2.1 Media composition and solutions for P. graminicola transformation

Media used to grow and reagents to transform cells are detailed in the tables below.

Table 5-1.YEPSL media composition (Brachmann et al. 2004)

| Raw Material | Concentration (g/L) |
|---|---|
| Yeast extract | 4 |
| Peptone | 4 |
| Sucrose | 20 |
| Deionised water | qs |

Table 5-2. Regeneration Agar composition (Brachmann et al. 2004)

| Raw Material | Concentration (g/L) |
|---|---|
| Sucrose | 20 |
| Peptone | 4 |
| Sorbitol | 182 |
| Yeast extract | 4 |
| Bacto agar | 15 |
| Deionised water | qs |

Table 5-3. Recipe for STC pH 7.5 solution (Brachmann et al. 2004)

| Raw Material | Concentration (g/L) |
|---|---|
| TRIS-HCl | 1,21 |
| $CaCl_2$ | 14,7 |
| Sorbitol | 182 |

Table 5-4. Recipe for SCS pH 5.8 solution (Brachmann et al. 2004)

| Raw Material | Concentration (g/L) |
|---|---|
| Sodium citrate | 5,875 |
| Sorbitol | 182 |

Table 5-5. Recipe for STC/PEG 4000 (Brachmann et al. 2004)

| Raw Material | Concentration |
|---|---|
| STC | 15 ml |
| PEG4000 | 10 g |

All solutions were autoclaved at 121°C at 15 lb of pressure.

### 5.3.3 Determination of antibiotic concentration for plate selection

In order to determine the best antibiotic concentration for *P. graminicola* plate selection, we prepared stock solutions at 1mg/mL for carboxin and 400 mg/ml for hygromycin. For these we prepared regeneration agar (RA) (Table 5-2) plates with 10, 20, 50, 100, 200, 500 and 1000 µL each of stock solution and recorded the growth on each plate. Both antibiotics worked successfully for transformation.

### 5.3.4 *P. graminicola* transformation

The transformation protocol was based on that of Brachmann et al. (2004) devised for *U. maydis* but with some modifications. *P. graminicola* CBS 10092 was incubated at 200 rpm overnight in 50 ml of YEPSL media (Table 5-1) in a 500 ml flask, at 30 °C. After 18 hours the culture was transferred to 200 ml of YEPSL liquid media in a 1000 ml flask. After about 3 h of further incubation at 30 °C, typically the cells reached the required optic density (value of 1.0) at 600 nm. 100 ml of culture was transferred into two falcon tubes (50 ml each) and centrifuged for 10 min at 1250xg (18 °C). The supernatant was poured off and the pellet was washed with 20 ml of SCS (Table 5-4) and centrifuged again. The supernatant was poured off and the pellet was resuspended in 1 ml of filter sterilized SCS containing 400 µL/ml VinoTaste Pro (Novozymes). Following digestion of the fungal cell wall for 3 h and 30 min at 37 °C, 2 ml of SCS was added, and the protoplasts were pelleted by centrifugation at 500xg (18 °C) for 2

min. To remove the remaining Novozyme, protoplasts were washed twice with 1 ml of SCS and once with 1 ml of STC (Table 5-3) and recovered by centrifugation as above. The pellet was resuspended in 0.5 ml of ice-cold STC. For transformation 50 µL of the protoplasts suspension was incubated for 10 min on ice with 5 µL of transforming DNA (at approximately 1 µg/µL) and 1 µL of a 15 mg/ml of heparin solution. After addition of 500 µL of STC/PEG400 (Table 5-5), the cell suspension was incubated for 15 min on ice. The whole transformation mix was then added to molten 10 ml of regeneration agar without the antibiotic, and plated on to selection media, which contained 10 mL of RA and 1 mg/mL carboxin. Colonies became visible after incubation for 3 days at 30 °C.

Successful homologous integration of the constructs was tested by colony PCR. For this, individual putative transformants were resuspended in 10 µL of PCR Biomix™ (Bioline) containing diagnostic primers for the specific targeted gene (Appendix ). The amplification took place with the following thermal profiles: denaturation at 95°C for 1 minute followed by 34 cycles of 15 seconds at 95°C, 15 seconds at 55 °C and 72°C for 2 minutes with a final elongation step for 5 minutes at 72°C. Successful clones were transferred to new 20 mL RA selection plates.

### 5.3.5 $^{1}$H NMR based semi-quantitative analysis of MEL production

We ran parallel micro-fermentations at GeneMill facility (Liverpool), using a Robolector XL (M2p labs). For each strain (WT and mutants) 150 µL of a 48 hours seed culture was used to inoculate 1500 µL of induction media (Table 2-1 and Table 2-2). We followed the same protocol as detailed in section 2.2.3.

The culture was incubated for three days at 30 °C at 120 rpm. 200 µL samples were taken every 24 hours. For cultures containing CRODAFAT, 96 mg was added every 24 hours and we implemented our optimised NMR protocol (Chapter 3) to process the samples and analyse the spectra.

#### 5.3.5.1 Statistical analysis for $^{1}$H NMR semi quantification

We followed the same protocol as detailed in section 3.3.7 for the spectral analysis and statistics.

### 5.3.6 Mutants morphology

Mutant and WT (as control) strains were grown in both YEPSL liquid media and plates (Table 5-1) and were incubated at 30°C for three days. For the liquid media, cultures a speed incubated with shaking (120 rpm).

For light microscopy, approximately 5 µL of a cell suspension was placed on a glass slide and cover with a cover slip and observed at 100X with oil immersion using an EVOS FL Cell Imaging System (Thermo Scientific).

## 5.4 RESULTS & DISCUSSION

### 5.4.1 Mutant analysis

In order to formally confirm the function of the MEL cluster we undertook to knock out *emt1*, which encodes a glucosyltransferase. We hypothesised that this would disrupt MEL production. To achieve this we developed *P. graminicola* transformation utilising antibiotic resistance cassettes for both carboxin and hygromycin resistance, driven by *U. maydis* arabinase promoters, $P_{crg1}$. We knocked out the *emt1* by implementing the method described by Brachmann and collaborators (2004) with some modifications (see 5.3.1) to transform *P. graminicola*'s strains

#### 5.4.1.1 emt1 deficient mutant

We knocked out the *emt1* gene from *P. graminicola* (PgEMT1) by introducing a gene marker coding for carboxin resistance flanked by the upstream and downstream genomics sequences of *emt1*, using homologous recombination, (

Figure 5-2). After three days of incubation the putative mutants had grown on the selection plates. To confirm the deletion of *emt1* individual transformants were subjected to PCR analysis (Figure 5-2) and a 350 bp fragment for the carboxin gene was produced for those colonies carrying the resistance marker. Additionally, only colonies which by PCRs amplified the entire sequence for resistance carboxyn gene and did not amplified for the WT *emt1* gene were classified as positive transformants.

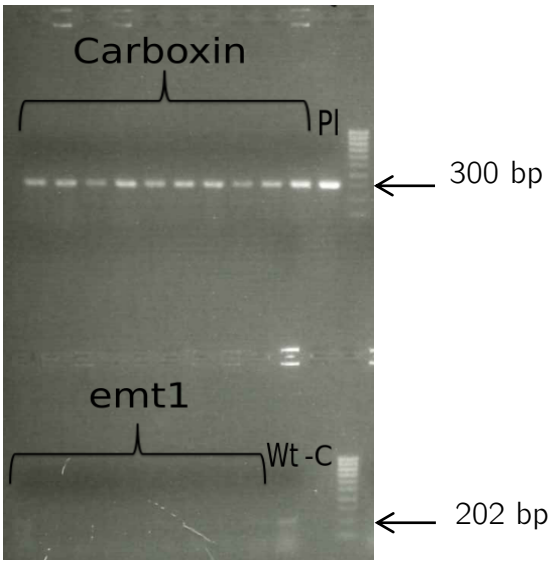



Figure 5-2. PCR confirmation for carboxin integration and *emt1* deletion. Upper panel corresponds to individual colonies positive for the carboxin gene, lower panel displaying same colonies negative for emt1. Pl= plasmid carrying carboxin resistance gene. WT= wild type DNA prep positive for emt1 gene, -C = water used as negative control.

The equivalent gene deletion was reported for *P. antarctica* T34 (Flagfeldt et al., 2009). They succeed in getting a full disruptive *emt1* but also obtained diploids, carrying one copy of the WT and one copy for the hygromycin resistance gene. To our knowledge the Δ*emt1* strain we produced corresponds to the first mutant of *P. graminicola* generated by gene targeting. Additionally, it confirms the utility of using carboxin and hygromycin resistance as the selection marker and that the arabinase *U. maydis* promoter for gene replacement and further genetic engineering in *P. graminicola.*

In *U. maydis*, deletion of *emt1* resulted in a subtle but distinct mutant morphology. Deletion did not interfere with the formation of conjugation hyphae but the cells appeared to stick to each other (Hewald et al. 2005). It was proposed that this aberrant morphology might be due to the strain's ability to produce two types of glycolipids (MELs and Ustilagic acid). In the case of *P. aphidis*, the Δ*emt1* strain displayed only the yeast form until 0.08% of MELs were added to the culture (Flagfeldt et al. 2009b). This suggests that MELs have a role in hyphal development.

Using light microscopy, we observed a change in *P. graminicola* morphology associated with deletion of *emt1* (Figure 5-3D and C respectively). There was a shift towards the yeast form accompanied by a dramatic reduction in cell size (WT average 10-15 μm, ΔPgEMT1 average 0.5-2 μm). Inclusion of approximately 5 μL of extracellular MEL-C (CRODA fraction, see Chapter 2) to the suspension culture appeared to have no effect in morphology of the mutant strain (48 hours). The growth morphology of WT and Δ*emt1* strain on plates also differed. The mutant showed a stronger yellow colouration and its appearance was glossy unlike the WT, which looked opaque (Figure 5-3, A and B respectively). These findings support the probable role of MELs in morphological development in *P. graminicola* and in spite of MELs being classed as secondary metabolites they appear to have a major role in normal growth and development in this species. Alternatively, it is possible that *emt1* has a role other than MEL production, which impacts of cell growth and morphology.

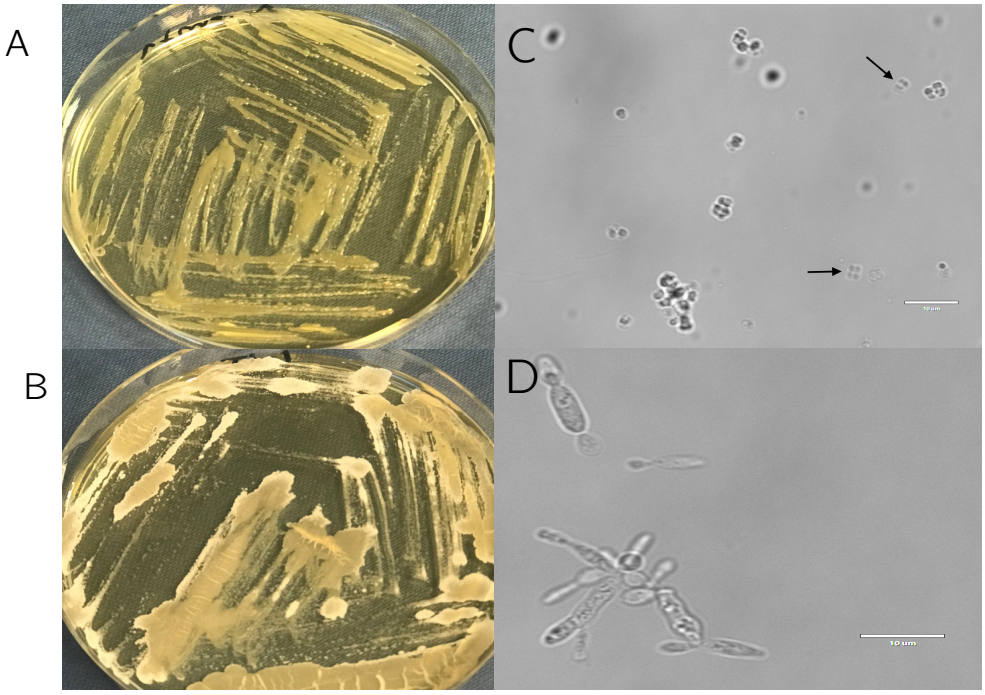



Figure 5-3. EMT1 WT and deficient mutant morphology. 48 hours cell culture grew in solid YEPSL media for A) ΔPgEMT1, B) WT. Magnification at 100X for C) ΔPgEMT1, D) WT. Scale bar= 10 um. Group of cells indicated with arrows.

#### 5.4.1.2 Semi-quantification of ΔPgEMT1 MEL production by 1H-NMR

EMT1 should be essential for MEL production. To confirm this, we compared the MEL production between the WT and the Δ*emt1* strains by semi-quantitative NMR analysis. We observed a significant reduction (t-test, $p < 0.05$) in the production of MEL-related metabolites in the mutant when compared to the WT (Figure 5-4). This confirms that emt1 is necessary for the production of MELs and this supports the functional annotation of *emt1* and the remaining genes in the MEL biosynthetic cluster.

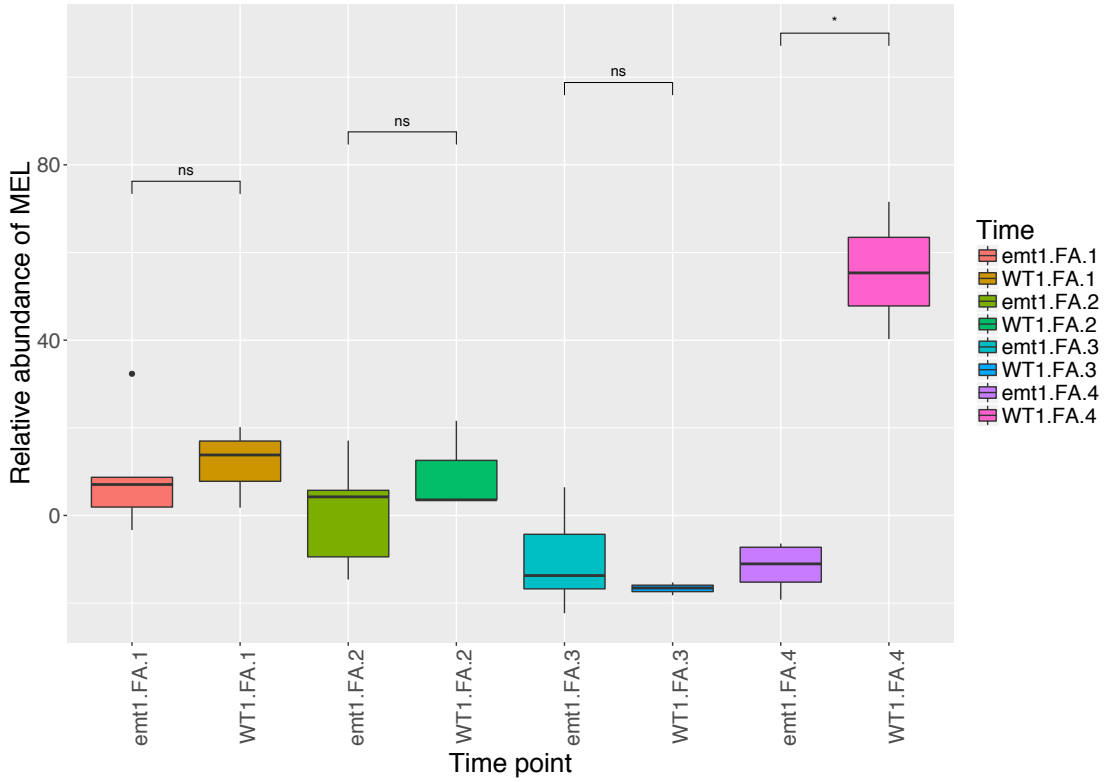



Figure 5-4. Relative abundance for MEL-related metabolites produced by WT and Δ*emt1* strains. micro-fermentations were conducted using the Biolector, with both WT and Δ*emt1* strains for 96h. Media samples were taken every 24 hours and analysed by NMR. The relative abundance of MEL related compounds was determined. A significant reduction ($p<0.05$) was observed for Δemt1 at the end of the fermentation. N=6.

### 5.4.2 Disruption of putative MEL regulators

In order to get a better understanding of potential regulatory mechanisms involved in MEL production we undertook to disrupt nitrogen regulation, which is suspected to be critically involved in MEL production (Hewald et al. 2005, 2006; Tollot et al. 2016) . Additionally, regulatory genes associated to mating and developmental functions in fungi have been linked to MEL production (Tollot et al. 2016), and therefore warrant investigation. To accomplish this, we identified the putative regulatory genes from the newly annotated genome and undertook targeted deletion utilising the same transformation protocol used to delete *emt1*.

#### 5.4.2.1 AREA protein deficient mutant

As MEL production requires nitrogen starvation, we decided to look for a homologue of positive-acting GATA transcription factor involved in nitrogen regulation and conserved in most fungi, AREA/Nit2 (Macios et al. 2012; Platt et al. 1996). AREA leads to the activation or co-activation of a large number of genes involved in the utilization of various nitrogen sources such as genes coding for catabolic enzymes and permeases (Caddick, 1994). Under nitrogen starvation conditions, AREA accumulates in the nucleus (Fraser et al., 2001; Todd et al., 2005). Therefore, we aimed to test whether the deletion of the orthologue in *P. graminicola* would affect MEL production, which is normally induced under nitrogen starvation. By amino acid comparison we found a putative *areA*-like gene in *P. graminicola,* designated *are1,* that showed a 46% similarity to the zinc-finger domain of *areA* from *A. nidulans* (Kudla et al. 1990).

A carboxin resistance *are1* deletion construct was produced and transformed into WT *P. graminicola*. Putative transformants were isolated and subjected to further analysis. Additionally, only colonies which by PCRs amplified the entire sequence for resistance carboxyn gene and did not amplified for the WT *areA* gene were classified as positive transformants. Unlike the situation in most fungi (eg *A. nidulans* (Macios et al. 2012) we were not able to fully delete the *are1* gene from *P. graminicola* (Figure 5-5). PCR analysis of individual transformants showed amplification for both *are1* $^{+}$ and the carboxin gene, suggesting the yeast strains obtained are heterozygous diploids. However, the possibility of heterologous integration cannot be excluded in this case.

This may indicate that *P. graminicola* is generally diploid, as is the case for *U. maydis* in its infective form and *P. antarctica T34*. It is noteworthy to mention that this mutant strain remained stable when plated on media without carboxin, indicating that it is not a heterokaryon or due to a wild type and mutant growing together.

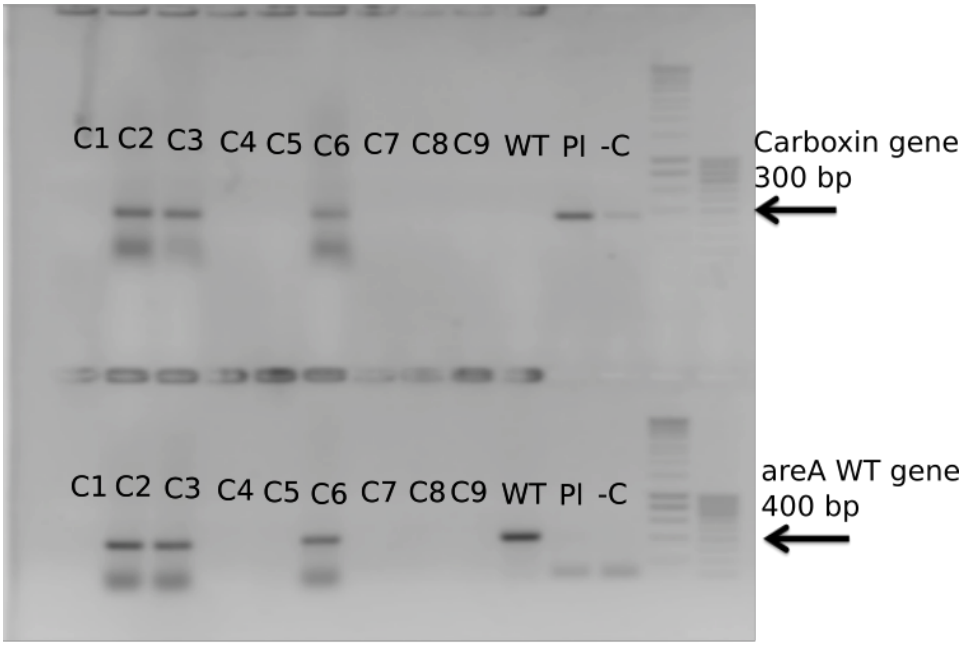



Figure 5-5. PCR confirmation for carboxin integration and *are1* deletion in *P. graminicola* transformants. Upper panel corresponds to PCR analysis of individual colonies for the carboxin gene, the lower panel displaying PCR analysis of the same colonies but for *areA1*. PI= plasmid carrying the carboxin gene. WT= wild type DNA prep, positive for *areA1* gene. –C= negative control with water in place of DNA. The three PCR positive transformants (C2, C3 and C6) are positive for both carboxin and *areA1,* indicative of them being heterozygous. Repeated experiments with different transformations gave the equivalent results with no homozygous *areA1* deletion strains being identified.

#### 5.4.2.2 ΔareA1 morphology

In pathogenic ascomycetes (*C. lindemuthianum*, *F. verticillioides*, and *F. oxysporum*) *areA/nit2* deficient strains showed reduced virulence on their respective hosts (Horst et al. 2012). In the case of filamentous fungi, studies showed that under nitrogen starvation conditions, the genes *area/nit2* enable the fungi to utilize complex nitrogen sources, but only few reports exist on these regulatory mechanisms in basidiomycete fungi (Horst et al. 2012; Macios et al. 2012; Platt et al. 1996). Although we could not fully delete this gene from *P. graminicola* the morphology of the mutant appears to have a yeast like form but differs from that of the *Δemt1* mutant (Figure 5-3). We observed signs of diploidy (two nuclei inside a cell) and potential cell division (Figure 5-6

C), which have been observed in *U. maydis* haploid cells switching to filamentous growth, when cultured in nitrogen starvation or low ammonium concentrations (Smith et al. 2003). Nevertheless, this does not explain the atypical morphology.

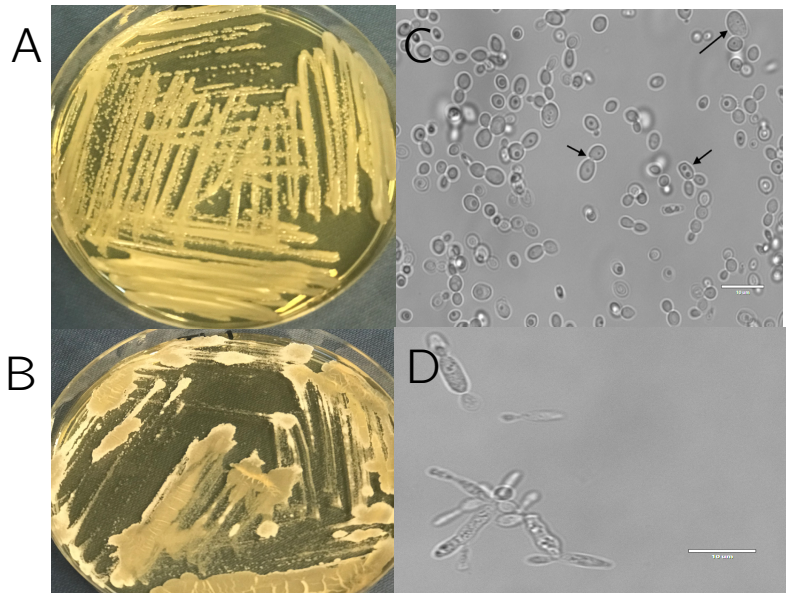



Figure 5-6. 48 hours cell culture for *areA1* deficient mutant and WT strains for *P. graminicola*. Growth in solid YEPSL media for A) ΔPgAREA, B) WT. Magnification at 100X for C) ΔPgAREA, D) WT. Scale bar= 10 um. Cells appearing to be undergoing cell division are indicated with arrows.

The heterozygous *areA1*$^{+}$/Δ*areA1* strain was assayed for MEL production and found not to differ much from the WT(Appendix 5-4) after 96 hours of fermentation. A complete *nit2/areA* deletion in the solopathogenic strain *U. maydis* SG200 (diploid strains that are pathogenic in the absence of mating) showed the same MEL composition under nitrogen starvation conditions when compared to its WT counterpart (Horst et al. 2012). However, as we were not able to test a strain homozygous for the *are1* deletion it is not possible to determine if the ARE1 transcription factor plays an important role, as was postulated on the basis of the prevalence of GATA motifs within the MEL gene cluster. Nevertheless, it should be noted that there are other GATA factors are

encoded by the *P. graminicola* genome, as in other fungi (i.e the fungal transcription factor regulatory middle homology region – TF MHR-, Pg6244).

### 5.4.3 Assessment of WOPR family members as potential MEL regulators

The WOPR protein family, comprises transcriptional regulators that bind DNA via their N-terminal WOPR box (Tollot et al. 2016). This consists of two domains: WOPRa and WOPRb, which are highly conserved. The presence of both domains is required to trigger the DNA binding activity (Lohse *et al*. 2010). In most of fungal genomes there are two paralogues, which fall into distinct clades (Caspari 1997; Kunitomo et al. 1995). It has been reported that members of this family positively regulate the production of secondary metabolites, are potentially involved in pathogenicity and regulation of sexual/asexual reproduction in phytopathogenic fungi (Michielse et al. 2011, Jonkers et al. 2012, Brown et al. 2014, Mirzadi et al. 2014, Okmen et al. 2014).

This family of transcription factors has been exclusively studied in *Ascomycetes*, with, to the best of our knowledge, one report for, analysing two members from *U. maydis*; *ros1 (UMAG_05853*) (regulator of sporogenesis) and *pac2 (UMAG_15096)*. Importantly, *ros1* has been shown to upregulate the MEL cluster (Tollot et al. 2016). Therefore, we looked for the corresponding *P. graminicola* orthologues and identified two genes Pg445 and Pg2545.

Comparative analysis of the respective proteins utilised WOPR family members that have been experimentally characterised in: *Fusarium verticillioides* Sge1 (FvSge1, W7MPI5), *Fusarium oxysporum f. sp. Lycopersici* Sge1 (FoSge1, AGA55574), *Verticillium dahlia* Sge1 (VdSge1, EGY16897), *Magnaporthe oryzae* GTI1 (MoGTI1, ELQ65940), *Fusarium graminearum* Fgp1 (FgFgp1, I1S5P3), *Botrytis cinerea* Reg1 (BcReg1, XP_001546736), *Candida albicans* Wor1 (CaWor1, Q5AP80), *Saccharomyces cerevisiae* Mit1 (ScMit1, P40002), *Schizosaccharomyces pombe* Gti1 (SpGti1, CAB61447), *Sporisorium reilianum*

SrWopr (CBQ70896) and *Cryptococcus neoformans* CnWopr (KIR63833). From a multiple amino acid alignment, the gene Pg445 from *P. graminicola* seems to share the two conserved domains from the WOPR box and the 15 amino acid residues corresponding to the R loop core DNA binding motif in Wor1 (Figure 5-7) (Tollot et al. 2016). This gene showed a high level of conservation with the orthologues from *U. maydis* (*ros1), S. pombe (gti1) S. cerevisiae (mit1*) with the encoded proteins having 63% ($e^{-0.00}$, 100% coverage), 55% ($2e^{-26}$, 20% coverage) and a 53% ($9e^{-19}$, 9% coverage) identity, respectively. Based on this we designated Pg445, *gti1.*


Figure 5-7. Section of amino acid sequence alignment of the WOPRa and WOPRb domains from *gti1* orthologous. Recognition loop (R loop) that recognizes the core DNA motif in Wor1 is indicated in black dotted box.

The putative protein encoded by Pg.2542 has high levels of identity with PAC2 from both *U. maydis* (66%, e-value 0.0, 84% coverage) and *S. pombe*'s (47%, e-value$^{4e-}$34, 19% coverage) and *S. cerevisiae Mit1p* (30%, e-value $1e^{-18}$, 23% coverage). Therefore, we named the Pg2542, *pac2.*

For the *PAC2* protein the motif associated with DNA recognition, WOPRa, Wor1 DNA binding motif is present (Figure 5-8) but the WOPRb box is missing.


Figure 5-8. Section of amino acid sequence alignment of the WOPRa segments from PAC2 orthologous. Recognition loop (R loop) that recognizes the core DNA motif in Wor1 is indicated in black dotted box. The WOPRb region is absent and therefore not displayed. UMAG_15069 (PAC2) included as best hit for PgPAC2.

### 5.4.4 GTI1 and PAC2 deficient mutants

In order to explore in more detail, the possible role of *P. graminicola gti1* and *pac2* in the regulation of MEL production, we proceed to create deletion mutants by gene replacement with *cbx* and succeeded in getting fully deleted strains, in both cases (Figure 5-9 and Figure 5-10). Additionally, only colonies which by PCRs amplified the entire sequence for resistance carboxyn gene and did not amplified for the WT *pac2* and *gti1* genes were considered as positive transformants.

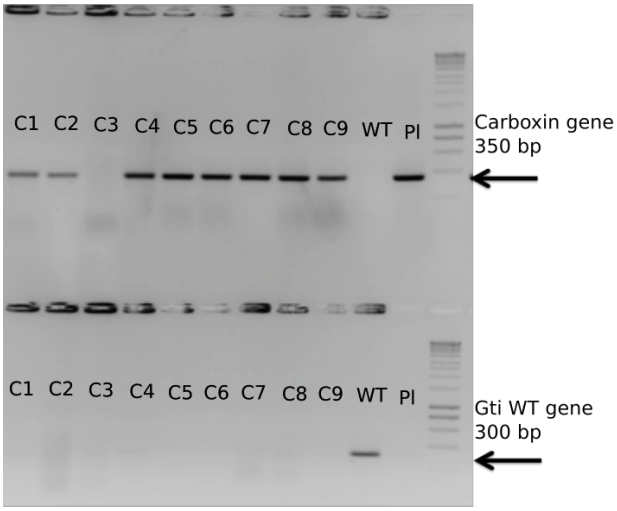



Figure 5-9. PCR confirmation for carboxin integration and deletion of gti1 for *P. graminicola* transformants. Upper panel corresponds to individual colonies positive for carboxin gene, lower panel displaying same colonies positive for gti1 WT gene. Pl= plasmid carrying the carboxin gene. WT= wild type DNA prep positive for gti1 gene. – C= water used as negative control.

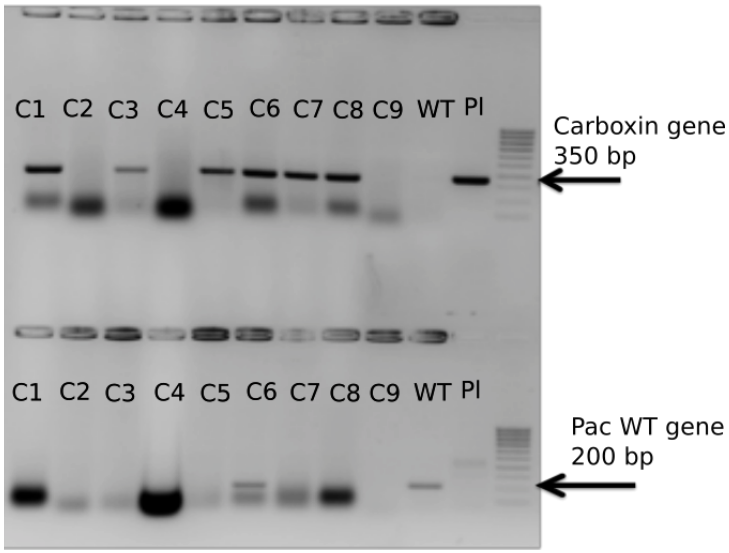



Figure 5-10. PCR confirmation for carboxin integration and deletion of *pac1* in *P. graminicola* transformants. Upper panel corresponds to individual colonies positive for carboxin gene, lower panel displaying same colonies positive for pac1 WT gene. Pl= plasmid carrying the carboxin gene. WT= wild type DNA prep positive for pac1 gene. – C= water used as negative control. C6 possible heterozygous or contamination.

### 5.4.5 Δ*gti1* and Δ*pac2* strain morphology

Both deficient Δ*gti1* and Δ*pac2* mutants strains differed in their morphology on YEPSL agar plates when compared to the WT strain (*Figure 5-11* A, B, C); *Δpac2* being the glossiest with a wet appearance common to both mutants. By light microscopy the differences between wild type and mutant cells were also noteworthy, both with respect to size and shape (*Figure 5-11* D, E, F). The WT presented primarily as elongated cells with an average size of 10-15 μm (*Figure 5-11* F) whereas *Δgti1* shifted towards a budding form with an average size of 0.5-2 μm (*Figure 5-11* D). We observed the same for the *Δpac2*, although most of its cells were within the range of 0.5-1 μm (*Figure 5-11* E). Following the observation by Flagfeldt and collaborators (2009b) in *P. aphidis*, where addition of 0.08% MEL to the media shifted the cells from budding to a yeast form, we added MEL-C enriched fraction to *Δgti1* and *Δpac2* fermentations and followed their morphology for 48 hours. We saw no change in morphology.

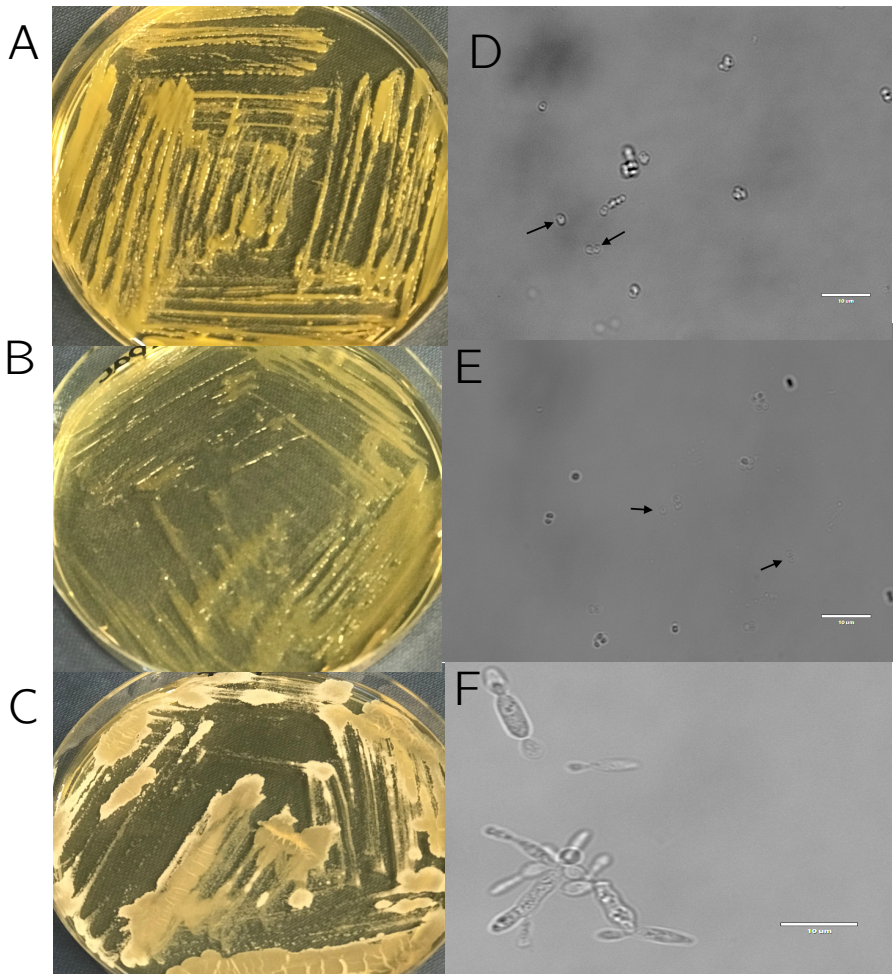



Figure 5-11. 48 hours cell culture for gt1 and pac2 deficient mutants and WT strains for *P. graminicola*. Growth in solid YEPSL media for A) ΔPgGTI1, B) ΔPgPAC2, C) WT. Magnification at 100X for D) ΔPgGTI1, E) ΔPgPAC2, F) WT. Scale bar= 10 um. Group of cells indicated with arrows.

The expressed morphology of the *Δpac2* mutant from *P. graminicola* met the observation reported in *Δpac2 U. maydis* strain, where a filament formation defect was identified due to the repression of hyphal growth (Elías-Villalobos, Fernández-Álvarez, and Ibeas 2011). Despite the molecular function of the protein being unclear, our observations being consistent with those of Elias-Villalobos *et al* (2011) supports the hypothesis of it having a role in filament formation and therefore, possibly infection.

#### 5.4.5.1 Semi-quantification of MEL production by Δgti1 and Δpac2 strains

We undertook fermentations and NMR analysis to assess the effect of the Δ*gti1* and Δ*pac2* alleles on MEL production. The data produced from Δ*gti1* strain, despite showing significant reduction on MEL production towards the end of the fermentation, is not entirely reliable (Figure 5-12). This, due to the massive decrease on MEL levels at 72 hours by the WT strain. We are aiming to repeat the experiment.

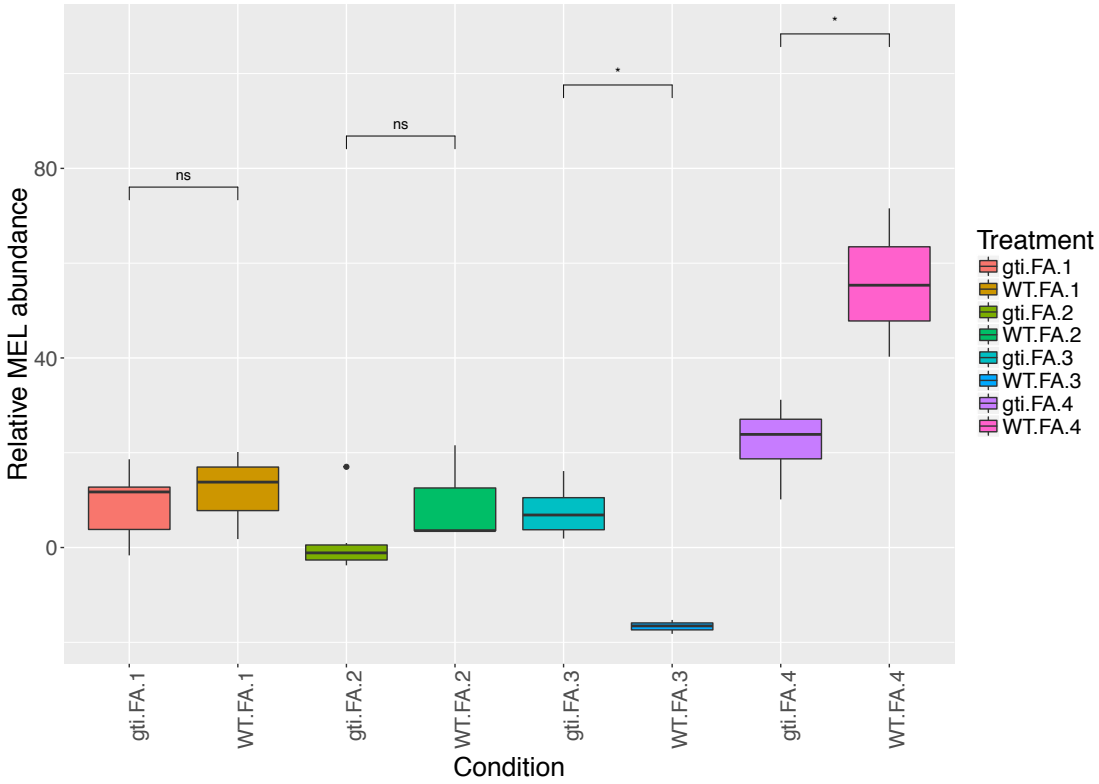



Figure 5-12. Relative abundance for MEL-related metabolites produced by WT and Δgti1 strains. Micro-fermentations were conducted using the Biolector, with both WT and Δgti1 strains for 96 h. media samples were taken every 24 hours and analysed by NMR. The relative abundance of MEL related compound was determined. A significant reduction (p< 0;05) was observed for Δgti1 at the end of the fermentation. N=6.

For the *Δpac2* strain, the effect of the mutation was the opposite to *Δgti1*, as MEL production was significantly enhanced (Figure 5-13). Nevertheless, as with *Δgti1,* we do not consider this data as entirely reliable due to the massive decrease of the WT. We repeated the experiment for this strain, but the results were yet

inconclusive. Additionally, the 96 hours sample for *ΔPgPAC2* could not be analysed by $^{1}$H NMR due to consistent failure on the QC spectra. This might be attributed to an excess on the MEL produced, affecting the signal.

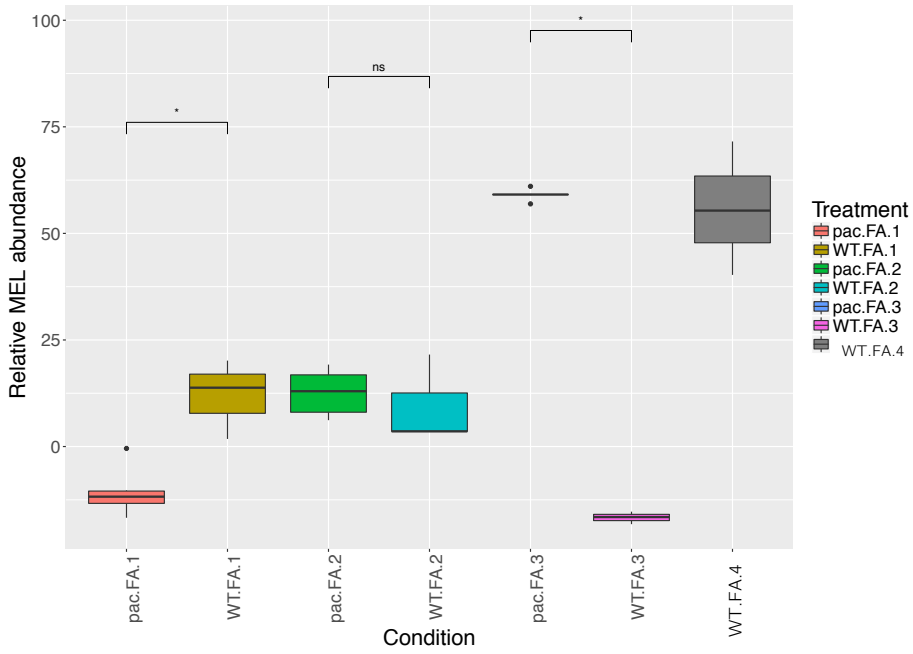



Figure 5-13. Relative abundance for MEL related metabolites produced by WT and Δpac2. Key: pac= mutant, WT= wild type. Time stamp corresponds to days (1 and 4 days). N=6. Significant increment ($p<0.05$) of MEL related compound is observed with Δ*pac2*.

## 5.5 CONCLUSIONS

We proved the successful transformation of *P. graminicola* strain and the use of *U. maydis* promoters, carrying selectable markers for carboxin resistance, for genetic engineering by fully deleting four out of the five genes we aimed.

Deletion of the glycosyltransferase encoding gene, *emt1*, disrupted MEL production consisted with it proposed role. Additionally, *emt1* disruption also changed the cell morphology, supporting the potential role of this gene in morphological switches between fungal and yeast form (Flagfeldt et al. 2009b).

Future analysis with this strain should include the supplementation of mannosylerythritol, in order to detect if the mutant strain is capable to form MEL. We found *gti1*, orthologues of which have been identified in many fungal species, including U. maydis *ros1*, may play a role in the regulation of MEL production. Deletion of *gti1*, lead to a significant reduction in MEL production suggesting it has a positive regulatory role. However, it is possible that this effect may be indirect. The GTI1 protein includes the WOPR box consistent with it being a functional DNA binding protein. Interestingly, orthologues of this gene have being implicated in fungal development, including sporulation, which would explain the changes we observed in morphology in the deleted strain. Potentially, it is possible that MELs may play a direct role in morphological development, but these two observations may not be directly linked.

Preliminary data demonstrated the deletion of a second putative transcription factor gene, *pac2*, also resulted in morphological changes and, more interestingly, in the enhanced production of MELs. As such both *pac2* and *gti1* may represent potential targets for industrial strain development.

In general, for the fully deleted *emt1, gti1* and *pac2* mutants, even though we obtained PCR products for an internal region of the resistance gene alongside with PCR products for the entire sequence and absence of PCR amplification for the WT gene, for future usage of this mutant verification with primers flanking different locations is important. For this, primers designed to amplify from either 5' or 3' end outside FR to the middle sequence of the resistance genes. Additionally, Southern blots could be implemented to evaluate possible multiple integration sites.

# 6. Concluding remarks and future work

This thesis describes the production of mannosylerythritol lipids in the fungus *P. graminicola*. We implemented metabolic profiling directed to MELs, *genomics* and *transcriptomics* in combination with biochemical and molecular genetic techniques, in order to accomplish our aims:

1. To sequence and assembly the genome of *P. graminicola*.
2. Obtain and utilise transcriptomic data to inform gene calling and identification of the MEL-C biosynthetic pathway.
3. Confirm the function of the putative MEL-C biosynthesis pathway by directed gene deletion and monitor expression of these and related genes involved in MEL production.

## 6.1. Results summary

The second chapter describes the sequencing, assembling and annotation of the *P. graminicola* genome from which we identified the genes comprising the MEL biosynthetic cluster and other genes potentially associated to pathogenicity. To fulfil this task, we utilised genomic and transcriptomic techniques. We implemented long-read sequencing with the PacBio platform to create a high quality assembled genome. Secondly, we generated and assembled the transcriptome by implementing a workflow of polyA selection for RNA depletion, short-sequencing pair-end reads (Illumina) and bioinformatics analysis.

In the third chapter we used *metabolomics* to identify and semi-quantify the production of MELs by *P. graminicola* in three different fermentative systems (fermenter, flasks and micro-fermentations) by $^{1}$H NMR. This led us to develop a protocol to process and systematically analyse samples from culture media regardless of the system used.

The fourth chapter, describes the characterisation of the MEL cluster genes and differentially expressed genes during producing and non-producing conditions by RNA-seq analysis. Additionally, we implemented qRT-PCR analysis, to monitor changes on the expression of genes from the MEL cluster. This, led us to conclude that the biosynthetic cluster genes are not co-regulated, nevertheless the presence of FA does induce expression.

In the fifth chapter we investigated the function of the MEL cluster and potential regulators of MEL production. For this, we created deletion mutants by using carboxin and higromycin resistance markers driven by arabinase promoters from *U. maydis*. This enable us to delete *emt1*, from which we confirmed a repression on the production of MELs. We also atempted to delete *areA/nit2*, but in this case we obtained heterozygotes carrying one copy of the WT allele and one deleted allele, replaced by the selection marker. Nevertheless, we identified two homologues of the WOPR family (Tollot et al., 2016; Elías-Villalobos, Fernández-Álvarez, & Ibeas, 2011). The deletion of *gti1* resulted in a drastic loss of the filamentous form in *P. graminicola* and a repression on MEL production. *pac2*, also displaying a distinct aberrant morphology, preliminary data suggests the mutant strain has enhanced production.

## 6.2. Contributions to the field

The first isolate of *P. graminicola* CBS 10092 was reported in 2007 in Moscow, Russia from herbaceous plants (Golubev, Sugita, & Golubev, 2007); one year later it was reported as being primarily a MEL-C producer (~85%), along with the

structural analysis (i.e fatty acid composition) and molecular properties (i.e surface tension) of the compound (Morita et al., 2008). Despite this, its genome had not been sequenced previously and its biology remains poorly characterised. Additionally, the current available genome sequences for the *Basidiomycota* group, were *P. graminicola* belongs, has fewer genomes than the group *Ascomycota*, meaning our genome impacts positively in future studies in a wide variety of fields. Among this, industrial biotechnology to dig deeper on the understanding of MEL production and regulation of *P. graminicola* and on the search to identify more clusters involved in the production of valuable secondary metabolites. The currently available genomes sequences for MEL producers comprised less than eight species, the majority being MEL-A producers (Alimadadi, Soudi, & Talebpour, 2018; Hewald et al., 2006; Konishi et al., 2008; Lorenz et al., 2014; Morita, Koike, et al., 2013; Saika, 2014; Saika et al., 2016). Therefore, the annotation of *P. graminicola* functions as a source for genomic information for future strain development towards MEL production.

By comparative genomics the function of unknown genes/proteins can be elucidated by inferring them from well described homologues (Wollenberg and Schirawski, 2014). We demonstrated *P. graminicola* harbours plenty of genes involved in pathogenicity on related strains, meaning the availability of this genome will allow more comparative studies in order to get a better understanding of the behaviour of effector genes.

This approach has been extensively used in the past five years and within the MEL producing community the identification of *Sporisorium* species has been accomplished (Alimadadi et al., 2018; Taniguti et al., 2015; Wollenberg & Schirawski, 2014). By following the same approach we also identified the presence of the effectors *cmu, hdp2, fox1, pep1* and the *pit* family in *P. graminicola*, which are known to be involved in plant defense, virulence suppression and in the establishment of the fungal biotrophic interaction (Lanver et al. 2017; Doehlemann et al. 2011) Taniguti et al., 2015). This finding suggests *P. graminicola* is

potentially a biotrophic pathogen and could be used as a vehicle to study and understand pathogenicity evolution and potential gene transference from related species, *U. maydis* and *S. reilianum*.

Noteworthy to mention we accomplished to construct deficient strains for *P. graminicola* by using *U. maydis* molecular machinery, this sets the precedent for future genetic engineering implementing the same transformation protocols

For biotechnological applications, the recovery of extracellular MELs characteristically involves the use of solvents and subsequent purification steps, hence our protocol for processing media samples by $^{1}$H NMR, involving one-step, is very convenient when systematisation and scaling is the target. Although this protocol only allows semi-quantification, we confirmed its utility in the efficient screening of multiple samples.

## 6.3. Trouble shooting: working with MELs

Initially we were provided with MEL standards by our industrial partner, CRODA, that derived from *P. aphidis*. Assessing these extracts by different analytical techniques (TLC, MS, NMR) revealed their low purity. Due to a lack of commercially available MEL standards we overcome this issue by comparing patterns of the different MEL forms (A, B, C and D) to our extracts. This approach allowed us to identify the presence and absence of MELs and to discriminate the doubly acetylated form, MEL-A, from the single acetylated form, MEL-B and MEL-C).

Furthermore, we aimed to implement mass spectrometry (MS) for the analysis of MELs but the time required to learn how to utilise the equipment, process and analyse the data in addition to the limited database at the GeneMill facility and the low purity standards limited our success. However, this is a priority for the future.

The metabolic dynamics for MEL production has been demonstrated to be highly variable, for as yet unknown reasons. Due to this variability the comparison between the three production systems (fermenter, flasks and micro-fermentations) required tailored optimisation, as the optimal conditions for one system had the opposite effect in another. This analysis the importance of factors such as agitation, FA concentration and aeration during MEL production.

## 6.4. Improvements and future work

Precise identification of specific MELs by any analytical technique requires pure standards for comparison, therefore it is imperative to purify MELs by both TLC and HPLC for its further analysis either by MS or NMR. As mentioned in previous chapters, the production of MELs has been reported mainly from expensive FA sources, such as soybean oil (Morita et al., 2007; Niu, Fan, Gu, Wu, & Chen, 2017) and olive oil (Jezierska et al., 2018; Morita, Fukuoka, Imura, & Kitamoto, 2013). Few studies report MEL production from post-refining waste, soaps (Dzięgielewska & Adamczak, 2013), waste frying oil (Fleurackers, 2006) and our study: using industrial waste FA. Therefore, in order to make MEL production financially viable and ecologically sustainable more studies implementing suples waste or unrefined sources of FA are required, ideally identifying the FA profile of the selected sources.

We aimed to monitor the expression of MELs by fusing gfp protein to the EMT1 glycosyltransferase, for this we used the carboxin and hygromycin resistance marker from *U. maydis* deletion cassette (Kämper, 2004). Nevertheless, we were not able to successfully visualise the GFP fusion. However, future studies monitoring the expression of the proteins over different time points and conditions, and monitoring the intracellular localisation, might provide valuable information to understand not only the kinetics of the protein but also the flux of production.

Additionally, these may provide valuable tools to select strains with high levels of expression, suing cell sorting.

## 6.5. Bottom line

MEL production in *basidiomycetes* has been studied for approximately 20 years with about 8 MEL producer genomes sequenced and available at the NCBI and yet the mechanisms for its regulation are not fully understood. This could be attributed to the vast majority of efforts directed to their biotechnological application leaving aside the valuable information of its biology. Our omic study contributes to elucidate not only key metabolite production but also the potential natural reasons behind its production. Nevertheless, some of our results shown interesting regulation, we are aware we need good repeatable experiments with transcriptomics and full metabolomic analysis to further apply systems biology to the system.

Additionally, implementation of metabolomics can also flux analysis to see where the components come from, identify key precursors and bottlenecks.

## REFERENCES


1. Adamczak M, Tomasik J, Płaszczyk M. 2004. "Application of oil refinery waste in the biosynthesis of glycolipids by yeast". *Bioresour Technol*. 95:15–8.
2. Adamczak, M and Włodzimierz B. 2000. "Influence of Medium Composition and Aeration on the Synthesis of Biosurfactants Produced by Candida Antarctica": 313–16.
3. Alberts B, Lewis J, Raff M, Roberts K, Walter P. 2002. "Molecular biology of the cell (4th ed." New York: Garland Science. ISBN 0-8153-3218-1.
4. Albrecht A, Rau U, Wagner F. 1996. "Initial steps of sophoroselipid biosynthesis by *Candida bombicola* ATCC 22214 grown on glucose". Appl Microbiol Biotechnol. 46:67-73
5. Alimadadi, N., M. R. Soudi, and Z. Talebpour. 2018. "Efficient Production of Tri-Acetylated Mono-Acylated Mannosylerythritol Lipids by *Sporisorium sp. aff. sorghi* SAM20." Journal of Applied Microbiology 124(2): 457–68.
6. Anders S and Wolfgang H. 2016. "Differential Expression of RNA-Seq Data at the Gene Level – the DESeq Package."
7. Anders S, Pyl PT and Huber W. 2015. "HTSeq - a Python framework to work with high-throughput sequencing data. Bioinformatics". 31:166–9.
8. Arnald A, Marsal S and JuliÃ A. 2015. "Analytical Methods in Untargeted Metabolomics: State of the Art in 2015." Frontiers in Bioengineering and Biotechnology 3(March): 1–20.
9. Arutchelvi, J. I., Bhaduri, S., Uppara, P. V., & Doble, M. 2008. "Mannosylerythritol lipids: a review". Journal of Industrial Microbiology & Biotechnology, 35(12), 1559–70.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig, J.T, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. "Gene ontology: tool for the unification of biology". The Gene Ontology Consortium. Nat. Genet., 25, 25–29.
11. Athenaki M, Gardeli C, Diamantopoulou P, Tchakouteu SS, Sarris D, Philippoussis A, Papanikolaou S. 2018. "Lipids from Yeasts and Fungi: Physiology, Production and Analytical Considerations." Journal of Applied Microbiology 124(2): 336–67.
12. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. "MEME SUITE: tools for motif discovery and searching". Nucleic Acids Res, 37(Web Server issue):W202–W208.
13. Bairoch A, Boeckmann B, Ferro S, Gasteiger E. 2004. "Swiss-Prot: juggling between evolution and stability". Brief Bioinformatics. 5:39–55.
14. Bamford NC, Snarr BD, Gravelat FN, Little DJ, Lee MJ, Zacharias CA, Chabot JC, Geller AM, Baptista SD, Baker P, Robinson H, Howell PL, Sheppard DC. 2015. "Sph3 is a glycoside hydrolase required for the biosynthesis of galactosaminogalactan in *Aspergillus fumigatus*". J. Biol. Chem. 290, 27438–27450.

15. Banat IM, Franzetti A, Gandolfi I, Bestetti G, Martinotti MG, Fracchia L, Smyth TJ, Marchant R. 2010. "Microbial Biosurfactants Production, Applications and Future Potential". Applied microbiology and biotechnology 87(2): 427–44.
16. Barth G, Gaillard C. 1997. "Physiology and genetics of the dimorphic fungus *Yarrowia lipolytica*". FEMS Microbiol Rev. 19:219-37.
17. Bartnicki-Garcia S. 1968. "Cell wall chemistry, morphogenesis, and taxonomy of fungi". Annu Rev Microbiol 22:87–108.
18. Bartoszewska, M., Opalinski, L., Veenhuis, M., and van der Klei, I.J. 2011."The significance of peroxisomes in secondary metabolite biosynthesis in filamentous fungi". Biotechnol Lett 33: 1921–1931.
19. Basse CW, Stumpferl S, Kahmann R. 2000. "Characterization of a *Ustilago maydis* gene specifically induced during the biotrophic phase: Evidence for negative as well as positive regulation". Mol. Cell. Biol. 20, 329–339.
20. Blencowe, B. J., Ahmad, S., and Lee, L. J. 2009. "Current-generation high throughput sequencing: deepening insights into mammalian transcriptomes". Genes & Development, 23(12), 1379–86
21. Brachmann, A, J. König, C. Julius, and M. Feldbrügge. 2004. "A Reverse genetic approach for generating gene replacement mutants in *Ustilago maydis*." Molecular Genetics and Genomics 272(2): 216–26.
22. Brakhage, Axel A. 2013. "Regulation of Fungal Secondary Metabolism". Nature Reviews Microbiology 11(1): 21–32.
23. Brefort, T., Tanaka, S., Neidig, N., Doehlemann, G., Vincon, V., and Kahmann, R. 2014. "Characterization of the Largest Effector Gene Cluster of *Ustilago maydis*". PLoS Pathogens 10(7).
24. Brown DW, Busman M, Proctor RH. 2014. "*Fusarium verticillioides* SGE1 is required for full virulence and regulates expression of protein effector and secondary metabolite biosynthetic genes". Molecular plant- microbe interactions: MPMI. 27(8):809–23.
25. Bruns TD, Vilgalys R, Barns SM, Gonzalez D, Hibbett DS, Lane DJ, Simon L, Stickel S, Szaro TM, Weisburg WG, Sogin ML. 1992. "Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences". Mol Phylogenet Evol 1:231–241.
26. Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2014. "Fast and Sensitive Protein Alignment Using DIAMOND." Nature Methods 12(1): 59–60.
27. Caddick, M.X., 1994. "Nitrogen metabolite repression". In: Martinelli, B.S., Kinghorn, J.R. (Eds.), Aspergillus: 50 Years On, Amsterdam–London–New York–Tokyo, 1994, pp. 323–353.
28. Calvo AM, Wilson RA, Bok JW, Keller NP. 2002. "Relationship between secondary metabolism and fungal development". Microbiology and molecular biology reviews: MMBR. 66(3):447–59, table of con- tents. PMID: 12208999; PubMed Central PMCID: PMC120793
29. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. 2009. "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics". Nucleic Acids Res. 37:D233–D238.
30. Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Yandell, M. 2008. "MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes". Genome Research, 18(1), 188–196.

31. Caspari T. 1997. "Onset of gluconate-H+ symport in *Schizosaccharomyces pombe* is regulated by the kinases Wis1 and Pka1, and requires the gti1+ gene product". J Cell Sci. 1997; 110 (Pt 20):2599–608. PMID: 9372449
32. Chidgeavadze Z, Beabealashvilli R. 1984. "2′, 3′-Dideoxy-3'aminonucleoside 5′-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases," Nucleic Acids Res. 12 (1984) 1671–1686
33. Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Korlach, J. 2013. "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data". Nature Methods, 10(6), 563–569.
34. Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., Wishart, D.S. and Xia, J. 2018. "MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis" . Nucl. Acids Res.
35. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Mortazavi, A. 2016. "A survey of best practices for RNA-seq data analysis". Genome Biology, 17(1), 13
36. Consortium, T. G. O. 2000. Gene ontologie: Tool for the unification of biology. Nature Genetics, 25(1), 25–29.
37. Costanzo, Samuel J. 1997. "Optimization of Mobile Phase Conditions for TLC Methods Used in Pharmaceutical Analyses." Journal of Chromatographic Science 35(4): 156–60.
38. Craig, Andrew et al. 2006. "Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets." Analytical Chemistry 78(7): 2262–67.
39. Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Yu, J. 2010. "A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing". Genomics, 96(5).
40. Cuomo CA, Birren BW. 2010. "The fungal genome initiative and lessons learned from genome sequencing". Methods Enzymol;470:833–55.
41. D., H. K., Brenner, S. E., and Dudoit, S. 2010. "Biases in Illumina transcriptome sequencing caused by random hexamer priming". Nucleic Acids Research, 38(12), 1–7.
42. Darling A, Bob M, Blattner F and Perna N. 2004. "Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements". 1394–1403.
43. Del Sorbo, G., H. Schoonbeek, and M. A. De Waard. 2000." Fungal trans- porters involved in efflux of natural toxic compounds and fungicides". Fungal Genet. Biol. 30:1–15.
44. Delcher AL, Salzberg SL, Phillippy AM. 2003. "Using MUMmer to identify similar regions in large sequence sets". Curr Protoc Bioinformatics Chapter 10: Unit 10 13.
45. Deml G, Anke T, Oberwinkler F, Gianetti BM, Steglich W. 1980. *Schizonellin* A and B, new glycolipids from *Schizonella melanogramma*. Phytochemistry 19:83–87.
46. Desai JD, Banat IM. 1997. "Microbial production of surfactants and their commercial potential". Microbiol Mol Biol Rev 61:47–64
47. Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. 2006. "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures". Application in $^{1}$H NMR metabonomics. Anal. Chem. 78, 4281–4290.

48. Djamei A, Schipper K, Rabe F, Ghosh A, Vincon V, Kahnt J, Osorio S, Tohge T, Fernie AR, Feussner I, Feussner K, Meinicke P, Stierhof YD, Schwarz H, Macek B, Mann M, Kahmann R. 2011. "Metabolic priming by a secreted fungal effector". Nature 478, 395–398.
49. Doehlemann, G., Reissmann, S., Aßmann, D., Fleckenstein, M., & Kahmann, R. 2011. "Two linked genes encoding a secreted effector and a membrane protein are essential for *Ustilago maydis*-induced tumour formation". Molecular Microbiology, 81(3), 751–766.
50. Dohren,H., Doonan,J., Driessen,A.J., Durek,P., Espeso,E.,Fekete,E., Flipphi,M., Estrada,C.G., Geysens,S., Goldman,G., de Groot,P.W., Hansen,K., Harris,S.D., Heinekamp,T., Helmstaedt,K.,Henrissat,B., Hofmann,G., Homan,T., Horio,T., Horiuchi,H., James,S., Jones,M., Karaffa,L., Karanyi,Z., Kato,M., Keller,N., Kelly,D.E., Kiel,J.A., Kim,J.M., van der Klei,I.J., Klis,F.M., Kovalchuk,A., Krasevec,N., Kubicek,C.P., Liu,B., Maccabe,A., Meyer,V., Mirabito,P., Miskei,M., Mos,M., Mullins,J., Nelson,D.R., Nielsen,J., Oakley,B.R.,Osmani,S.A., Pakula,T., Paszewski,A., Paulsen,I., Pilsyk,S., Pocsi,I., Punt,P.J., Ram,A.F., Ren,Q., Robellet,X., Robson,G., Seiboth,B., van Solingen,P., Specht,T., Sun,J., Taheri-Talesh,N., Takeshita,N., Ussery,D., vanKuyk,P.A.,Visser,H., van de Vondervoort,P.J., de Vries,R.P., Walton,J.,Xiang,X., Xiong,Y., Zeng,A.P., Brandt,B.W., Cornell,M.J., van den Hondel,C.A., Visser,J., Oliver,S.G. and Turner,G. 2008. "The 2008 update of the *Aspergillus nidulans* genome annotation: a community effort". Fungal Genet. Biol. 46 (SUPPL 1), S2-S13.
51. Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, and Botstein D 2002. Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*". Proc Natl Acad Sci U S A. 2002;99:16144–9.
52. E.L. van Dijk, H. Auger, Y. Jaszczyszyn, C. Thermes. 2014. "Ten years of next-generation sequencing technology". Trends Genet. 30.
53. Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." Nucleic Acids Research 32(5): 1792–97.
54. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X.X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C.C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. "Real-time DNA sequencing from single polymerase molecules". Science 323, 133–138.
55. Ekblom, Robert, and Jochen B W Wolf. 2014. "A Field Guide to Whole-Genome Sequencing, Assembly and Annotation." Evolutionary Applications 7(9): 1026–42.
56. Elías-Villalobos A, Fernández-Álvarez A and Ibeas J. 2011. "The General Transcriptional Repressor Tup1 Is Required for Dimorphism and Virulence in a Fungal Plant Pathogen." PLoS Pathogens 7(9): 27–33.
57. Emms D and Kelly S. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." Genome Biology 16(1): 1–14.

58. Emms D and Kelly S. 2018. "STAG: Species Tree Inference from All Genes." bioRxiv.
59. Ewelina D, and Adamczak M. 2013. "Evaluation of Waste Products in the Synthesis of Surfactants by Yeasts". Chemical Papers 67(9): 1113–22.
60. Fan LL, Dong YC, Fan YF, Zhang J, Chen QH. 2014. "Production and Identification of Mannosylerythritol Lipid-A Homologs from the *Ustilaginomycetous* yeast *Pseudozyma aphidis* ZJUDM34." Carbohydrate Research 392: 1–6.
61. Faria, N. T., Santos, M. V., Fernandes, P., Fonseca, L. L., Fonseca, C., and Ferreira, F. C. 2014." Production of glycolipid biosurfactants, mannosylerythritol lipids, from pentoses and d-glucose/d-xylose mixtures by *Pseudozyma* yeast strains". Process Biochemistry, 49(11), 1790–1799
62. Flagfeldt B, Siewers V, Huang L and Nielsen J. 2009a. "Characterization of Chromosomal Integration Sites for Heterologous Gene Expression in *Saccharomyces cerevisiae*." Yeast (Chichester, England) 26(10): 545–51.
63. Flusberg B; Webster D, Lee Jessa, Travers K, Olivares E, Clark T, Korlach J and Tuner S. 2010. "Direct detection of DNA methylation during single-molecule, real-time sequencing". Nat. Methods 7, 461–465 (2010). 93.
64. Fraser, J.A., Davis, M.A., Hynes, M.J., 2001. "The formamidase gene of *Aspergillus nidulans*: regulation by nitrogen metabolite repression and transcriptional interference by an overlapping upstream gene". Genetics 157, 119–131.
65. Freitag J, Julia A, Uwe L, Thorsten S, Domenica M, Michael B and Björn S. 2014. "Peroxisomes Contribute to Biosynthesis of Extracellular Glycolipids in Fungi." Molecular Microbiology 93(June): 24–36.
66. Fukuoka T, Morita T, Konishi M, Imura T, Kitamoto D .2007. "Characterisation of new types of mannosylerythritol lipids as bio- surfactants produced from soybean oil by a *Basidiomycetes* yeast, *Pseudozyma shanxiensis*". J Oleo Sci 56:435–442
67. Fukuoka T, Morita T, Konishi M, Imura T, Kitamoto D. 2008. "A *Basidiomycetous* Yeast, *Pseudozyma tsukubaensis*, Efficiently Produces a Novel Glycolipid Biosurfactant. The Identification of a New Diastereomer of Mannosylerythritol Lipid-B." Carbohydrate Research 343(3): 555–60.
68. Galagan JE, Henn MR, Ma L-J, Cuomo C, Birren B. 2005. "Genomics of the fungal kingdom: insights into eukaryotic biology". Genome Res. 15:1620–31.
69. Gibson DG, Smith HO, Hutchison CA 3rd, Venter JC, Merryman C. 2010. "Chemical synthesis of the mouse mitochondrial genome". Nat. Methods 7, 901–903.
70. Golubev, W., Sugita, T and Golubev, N. 2007. "An *Ustilaginomycetous* yeast, *Pseudozyma graminicola* sp. nov., isolated from the leaves of pasture plants". Mycoscience, 48(1), 29–33
71. Goodwin S, McPherson J and McCombie R. 2016. "Coming of Age: Ten Years of next-Generation Sequencing Technologies." Nature Reviews Genetics 17(6): 333–51.
72. Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Conesa, A. 2008. "High-throughput functional annotation and data mining with the Blast2GO suite". Nucleic Acids Research, 36(10), 3420–3435.

73. Grigoriev I, Cullen D, Goodwin S, Hibbett D, Jeffries T, Kubicek C, Kuske C, Magnuson J, Martin F, Spatafora J, Tsang A and Baker S. 2011. "Fueling the future with fungal genomics". Mycology. 2(3):192–209.
74. Guigó, R., Agarwal, P., Abril, J. F., Burset, M., and Fickett, J. W. 2000. "An assessment of gene prediction accuracy in large DNA sequences". Genome Research, 10(10), 1631–42.
75. Günther M, Grumaz C, Lorenz S, Stevens P, Lindemann E, Hirth T, Sohn K, Zibek S and Rupp S. 2015. "The transcriptomic profile of *Pseudozyma aphidis* during production of mannosylerythritol lipids". Applied Microbiology and Biotechnology, 99(3), 1375–1388.
76. Guo, Y., Cordes, K.R., Farese Jr., R.V. and Walther, T.C. 2009. "Lipid Droplets at a Glance". Journal of Cell Science, 122, 749-752.
77. Gupta, P.K., 2008. "Single-molecule DNA sequencing technologies for future genomics research". Trends Biotechnol. 26, 602–611.
78. Halsall, J. R., M. J. Milner, and L. A. Casselton. 2000. "Three subfamilies of pheromone and receptor genes generate multiple B mating specificities in the mushroom *Coprinus cinereus*". Genetics 154:1115–1123.
79. Hayden E. 2015."Pint-sized DNA sequencer impresses first users". Nature 521: 15–16.
80. Heather, James M, and Benjamin Chain. 2016. "Genomics The Sequence of Sequencers : The History of Sequencing DNA." Genomics 107(1): 1–8.
81. Hedges, S. Blair. 2002. "The Origin and Evolution of Model Organisms." Nature Reviews Genetics 3(11): 838–49.
82. Helga D, Ozçelik I, Hofmann G, and Nielsen J. 2008. "Analysis of *Aspergillus nidulans* Metabolism at the Genome-Scale." BMC genomics 9: 163.
83. Hemetsberger C, Herrberger C, Zechmann B, Hillmer M and Doehlemann G. 2012. "The *Ustilago maydis* effector Pep1 suppresses plant immunity by inhibition of host peroxidase activity. PLoS Pathogens. 8, e1002684.
84. Hewald S, Josephs K, Bölker M and Bo M. 2005. "Genetic Analysis of Biosurfactant Production in *Ustilago maydis*." Applied and environmental microbiology 71(6): 3033–40.
85. Hewald S, Uwe L, Scherer M, Mohamed M, Jörg K, Bölker M. 2006. "Identification of a Gene Cluster for Biosynthesis of Mannosylerythritol Lipids in the Basidiomycetous Fungus *Ustilago maydis*." Applied and environmental microbiology 72(8): 5469–77.
86. Hibbett DS, Binder M, Bischoff JF et al. 2007. "A higher-level phylogenetic classification of the Fungi". Mycol Res 111: 509–547.
87. Hibbett DS. 2006. "A phylogenetic overview of the *Agaricomycotina*". Mycologia 98: 917–925.
88. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. 2016. "BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 1;32(5):767-9.
89. Hoffmeister, D., and Keller, N.P. 2007. "Natural products of filamentous fungi: enzymes, genes, and their regulation". Nat Prod Rep 24: 393–416.
90. Holmberg K.: 2001. "Natural surfactants". Curr. Opin. Colloid Interface Sci., 6, 148-159.

91. Horst J, Zeh C, Saur A, Sonnewald S, Sonnewald U, Voll M. 2012. "The *Ustilago maydis* Nit2 Homolog Regulates Nitrogen Utilization and Is Required for Efficient Induction of Filamentous Growth." Eukaryotic Cell 11(3): 368–80.
92. Isoda H, Kitamoto D, Shinmoto H, Matsumura M and Nakahara T. 1997. "Microbial Extracellular Glycolipid Induction of Differentiation and Inhibition of the Protein Kinase C Activity of Human Promyelocytic Leukemia Cell Line HL60." Bioscience, biotechnology, and biochemistry 61(4): 609–14.
93. Jang-il S and Jin-Wu N. 2016. "The present and future of de novo whole-genome assembly." Briefings in Bioinformatics (June): bbw096.
94. Jezierska S, Silke C and Van Bogaert I. 2018. "Yeast Glycolipid Biosurfactants." FEBS Letters 592(8): 1312–29.
95. Jones P; Binns D, Chang Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn A, Sangrador-Vegas A, Scheremetjew M, Yong S, Lopez R and Hunter S. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." Bioinformatics 30(9): 1236–40.
96. Jonkers W, Dong Y, Broz K, Kistler HC. 2002. "The Wor1-like protein Fgp1 regulates pathogenicity, toxin synthesis and reproduction in the phytopathogenic fungus *Fusarium graminearum*". PLoS pathogens. 2012; 8(5):e1002724.
97. Josephs K, Bölker M, and Bo, M. 2005. "Genetic analysis of Biosurfactant production in Ustilago maydis. Applied and Environmental Microbiology, 71(6), 3033–3040.
98. Kämper, J. 2004. "A PCR-Based System for highly efficient generation of gene replacement mutants in *Ustilago Maydis*." Molecular Genetics and Genomics 271(1): 103–10.
99. Kämper, J., Kahmann, R., Bölker, M., Ma, L.J., Brefort, T., Saville, B.J., Banuett, F., Kronstad, J.W., Gold, S.E., Müller, O., Perlin, M.H., Wösten, H.A.B., De Vries, R., Ruiz-Herrera, J., Reynaga-Peña, C.G., Snetselaar, K., McCann, M., Pérez-Martín, J., Feldbrügge, M., Basse, C.W., Steinberg, G., Ibeas, J.I., Holloman, W., Guzman, P., Farman, M., Stajich, J.E., Sentandreu, R., González-Prieto, J.M., Kennell, J.C., Molina, L., Schirawski, J., Mendoza-Mendoza, A., Greilinger, D., Münch, K., Rössel, N., Scherer, M., Vraněs, M., Ladendorf, O., Vincon, V., Fuchs, U., Sandrock, B., Meng, S., Ho, E.C.H., Cahill, M.J., Boyce, K.J., Klose, J., Klosterman, S.J., Deelstra, H.J., Ortiz-Castellanos, L., Li, W., Sanchez-Alonso, P., Schreier, P.H., Häuser-Hahn, I., Vaupel, M., Koopmann, E., Friedrich, G., Voss, H., Schlüter, T., Margolis, J., Platt, D., Swimmer, C., Gnirke, A., Chen, F., Vysotskaia, V., Mannhaupt, G., Güldener, U., Münsterkötter, M., Haase, D., Oesterheld, M., Mewes, H.W., Mauceli, E.W., DeCaprio, D., Wade, C.M., Butler, J., Young, S., Jaffe, D.B., Calvo, S., Nusbaum, C., Galagan, J., Birren, B.W.,. 2006. "Insights from the Genome of the Biotrophic Fungal Plant Pathogen *Ustilago maydis*." Nature 444(7115): 97–101.
100. Keibler, E., and Brent, M. R. 2003. "Eval: a software package for analysis of genome annotations". BMC Bioinformatics, 4(1), 50.
101. Kenneth J. Livak and Thomas D. Schmittgen. 2001. "Analysis of Relative Gene Expression Data Using Real- Time Quantitative PCR and the $2^{-C_T}$ Method". METHODS **25,** 402–408.

102. Kim H. S. Yoon B. D, Choung D. H, Oh H, Katsuragi T and Tani Y. 1999. "Characterization of a Biosurfactant, Mannosylerythritol Lipid Produced from *Candida sp*. SY16." Applied Microbiology and Biotechnology 52(5): 713–21.
103. Kitamoto D, Hiroko I and Tadaatsu N. 2002. "Functions and Potential Applications of Glycolipid Biosurfactants--from Energy-Saving Materials to Gene Delivery Carriers." Journal of bioscience and bioengineering 94(3): 187–201.
104. Kitamoto D, Ikegami T, Tiemi S, Sasaki G, Takeyama A, Idemoto Y; Nobuyaki K, Yanagishita H. 2001. "Microbial Conversion of N-Alkanes into glycolipid biosurfactants, nanosylerytrol lipids, by Psedozyma (Candida antarctica)." Biotechnol letters 23: 1709–14.
105. Kitamoto D, Kazuaki H, Tadaatsu N and Takeshi T. 1990. "Production of Mannosylerythritol Lipids by *Candida antarctica* from Vegetable Oils." Agric. Biol. Chem. 54(1): 37–40.
106. Kitamoto D, Morita T, Fukuoka T, Konishi M and Imura, T. 2009. "Self-assembling properties of glycolipid biosurfactants and their potential applications". Curr. Opin. Colloid Interface Sci. 14:315–328.
107. Kitamoto D, Nemoto, T and Yanagishita, H.1990."Fatty acid metabolism of mannosylerythritol lipids as Biosurfactants produced by *C. antarctica".* J. Jpn. Oil Chem. Soc. Vol. 42. No. 5:346-358.
108. Kitamoto D, Shyunichi A, Chieko H and Takeshi T. 1990. "Extracellular Accumulation of Mannosylerythritol Lipids by a strain of Candida antarctica." Agricultural and Biological Chemistry 54(1): 31–36.
109. Konishi M and Motoki M. 2017. "Selective production of deacetylated Mannosylerythritol lipid, MEL-D, by acetyltransferase disruption mutant of Pseudozyma hubeiensis." Journal of Bioscience and Bioengineering 125(1).
110. Konishi M, Imura T, Fukuoka T, Morita T, Kitamoto D .2007. "A yeast glycolipid biosurfactant, mannosylerythritol lipid, shows high binding affinity towards lectins on a self-assembled monolayer system". Biotechnol Lett 29:473–480.
111. Konishi M, Morita T, Fukuoka T, Imura T, Kakugawa K, Kitamoto D. 2008. "Efficient production of Mannosylerythritol lipids with high hydrophilicity by Pseudozyma hubeiensis KM-59." Applied Microbiology and Biotechnology 78(1): 37–46.
112. Konishi M, Tokuma F, Takahiko N, Tomotake M, Tomohiro I, Dai K, Yuji H. 2018. "Selective production of deacetylated Mannosylerythritol lipid, MEL-D, by acetyltransferase disruption mutant of *Pseudozyma hubeiensis*." Journal of Bioscience and Bioengineering 125(1): 105–10.
113. Konishi M, Tokuma F, Takahiko N, Tomotake; M, Tomohiro I, Dai K, Yuji H. 2010. "Biosurfactant-producing yeast isolated from *Calyptogena soyoae* (Deep-sea cold-seep clam) in the deep sea." Journal of Bioscience and Bioengineering 110(2): 169–75.
114. Konishi M, Yoshida Y, Ikarashi M and Horiuchi J. 2015. "Efficient and simple electro-transformation of intact cells for the Basidiomycetous fungus *Pseudozyma hubeiensis*." Biotechnology Letters 37(8): 1679–85.
115. Konishi, M., Hatada, Y., and Horiuchi, J. 2013. "Draft Genome sequence of the Basidiomycetous yeast-like fungus *Pseudozyma hubeiensis* SY62, which

produces an abundant amount of the biosurfactant Mannosylerythritol Lipids". Genome Announcements, 1(4), 13–14

116. Koren, S., and Phillippy, A. M. 2015. "One chromosome, one contig: Complete microbial genomes from long-read sequencing and assembly. Current Opinion in Microbiology, 23, 110–120.
117. Korf I, Yandell M and Bedell J. 2003. "An essential guide to the Basic Local Alignment Search Tool". O'Reilly. First Edition. United States of America.
118. Koszul R, Caburet S, Dujon B and Fischer G. 2004. "Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments". EMBO J. 23:234–43.
119. Kudla B, Caddick M, Langdon T, Martinez-Rossi N, Bennett C, Sibley S and Arst H. 1990. "The regulatory gene AreA mediating nitrogen metabolite repression in Aspergillus nidulans. Mutations affecting specificity of gene activation alter a loop residue of a putative zinc finger." The EMBO journal 9(5): 1355–64.
120. Kunitomo H, Sugimoto A, Wilkinson CR, Yamamoto M. 1995. Schizosaccharomyces pombe pac2+ controls the onset of sexual development via a pathway independent of the cAMP cascade. Curr Genet.28(1):32–8. PMID: 8536311.
121. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S. L. 2004." Versatile and open software for comparing large genomes". Genome Biology, 5(2), R12
122. Lanver, D., Tollot, M., Schweizer, G., Lo Presti, L., Reissmann, S., Ma, L.-S., Kahmann, R. 2017. "Ustilago maydis effectors and their impact on virulence". Nature Reviews Microbiology, 15(May), 409–421.
123. Lee, H., Gurtowski, J., and Yoo, S. 2014. "Error correction and assembly complexity of single molecule sequencing reads". bioRxiv, 1–17.
124. Li, L., Stoeckert, C. J. J., and Roos, D. S. 2003. "OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes" Genome Research, 13(9), 2178–2189.
125. Liu, Y.M., Zhang, C.Y., Shen, X.P., Zhang, X.L., Cichello, S., Guan, H.B. and Liu, P.S. 2013. "Microorganism Lipid Droplets and Biofuel Development". BMB Reports, 46, 575-581.
126. Livak K and Schmittgen T. 2001. "Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method." Methods 25(4): 402–8.
127. Lohse MB, Zordan RE, Cain CW and Johnson AD. 2010. "Distinct class of DNA-binding domains is exemplified by a master regulator of phenotypic switching in *Candida albicans*". Proc Natl Acad 107 (32):14105–10.
128. Lomsadze, A., Burns, P. D., and Borodovsky, M. 2014. "Integration of mapped RNA-Seq-reads into automatic training of eukaryotic gene finding algorithm". Nucleic Acids Research, 42(15), 1–8.
129. Lorenz S, Guenther M, Grumaz C, Rupp S, Zibek S, Sohn K. 2014. "Genome sequence of the Basidiomycetous fungus Pseudozyma aphidis DSM70725 , an Efficient Producer of Biosurfactant Mannosylerythritol." 2(1): 13–14.

130. Lössl P, van de Waterbeemd M and Heck A. 2016. "The diverse and expanding role of Mass Spectrometry in structural and molecular biology." The EMBO Journal 35(24): 2634–57.
131. Love, M. I., Simon A and Wolfgang H. 2014. "Genome biology differential analysis of Count Data - the DESeq2 Package". Genome Biology. 15 (12):550.
132. Macios M, Caddick M, Weglenski P, Scazzocchio C and Agnieszka Dzikowska A. 2012. "The GATA factors AREA and AREB together with the co-repressor NMRA, negatively regulate arginine catabolism in Aspergillus nidulans in response to nitrogen and carbon source." Fungal genetics and biology : FG & B 49(3): 189–98.
133. Malik A, Firoz A, Jha V and Ahmad S. 2010. "PROCARB: A database of known and modelled carbohydrate-binding protein structures with sequence-based prediction tools." Advances in Bioinformatics 2010.
134. Mardis, E. R. 2008. "Next-Generation DNA sequencing methods". Annual Review of Genomics and Human Genetics, 9(1), 387–402.
135. Marion, Dominique. 2013. "An Introduction to Biological NMR Spectroscopy." Molecular & Cellular Proteomics 12(11): 3006–25.
136. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J, Helfenbein KG, Ramaiya P, Detter JC, Larimer F, Coutinho PM, Henrissat B, Berka R, Cullen D, Rokhsar D. 2004. "Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78". Nat Biotechnol; 22:695–700.
137. Michielse C, Becker M, Heller J, Moraga J, Collado IG, Tudzynski P. 2011. "The Botrytis cinerea Reg1 protein, a putative transcriptional regulator, is required for pathogenicity, conidiogenesis, and the production of secondary metabolites". Molecular plant-microbe interactions: MPMI; 24(9):1074–85.
138. Michielse C, van Wijk R, Reijnen L, Manders E, Boas S, Olivain C, Alabouvette C, Rep M. 2009. "The nuclear protein Sge1 of Fusarium oxysporum is required for parasitic growth". PLoS pathogens. 5(10):e1000637.
139. Mirzadi A, Mehrabi R, Robert O, Agah I, Boeren S, Schuster M, Steinberg G, de Vit P and Kema G. 2014. "Molecular characterization and functional analyses of ZtWor1, a transcriptional regulator of the fungal wheat pathogen Zymoseptoria tritici:. Mol Plant Pathol. 2014; 15(4):394–405.
140. Morita T, Fukuoka T, Imura T, Kitamoto D. .2013a. "Accumulation of cellobiose lipids under nitrogen-limiting conditions by two Ustilaginomycetous yeasts, Pseudozyma aphidis and Pseudozyma hubeiensis." FEMS Yeast Research 13: 44–49.
141. Morita T, Habe H, Fukuoka T, Imura T, Kitamoto D. 2007. "Convenient transformation of anamorphic Basidiomycetous yeasts belonging to genus Pseudozyma induced by electroporation." Journal of bioscience and bioengineering 104(6): 517–20.
142. Morita T, Koike H, Hagiwara H, Ito E, Machida M, Sato S, Habe H, Kitamoto D. 2014b. "Genome and transcriptome analysis of the Basidiomycetous yeast Pseudozyma antarctica producing extracellular Glycolipids, Mannosylerythritol Lipids." PLoS ONE 9(2).
143. Morita T, Koike H, Koyama Y, Hagiwara H, Ito E, Fukuoka T, Imura T, Machida M, Kitamoto D. 2013. "Genome Sequence of the Basidiomycetous

yeast Pseudozyma antarctica T-34, a producer of the glycolipid biosurfactants." 1(2): 1–2.

144. Morita T, Konishi M, Fukuoka T, Imura T, Kitamoto D .2006. "Discovery of Pseudozyma rugulosa NBRC 10877 as a novel producer of the glycolipid biosurfactants, mannosylerythritol lipids, based on rDNA sequence. Appl Microbiol Biotechnol 73:305– 313.

145. Morita T, Konishi M, Fukuoka T, Imura T, Kitamoto HK, Kitamoto D. 2007. "Characterization of the genus Pseudozyma by the formation of glycolipid biosurfactants, mannosylerythritol lipids". FEMS Yeast Res 7:286–292.

146. Morita T, Konishi M, Fukuoka T, Imura T, Yamamoto S, Kitagawa M, Sogabe A, Kitamoto D. 2008. "Identification of Pseudozyma graminicola CBS 10092 as a producer of glycolipid biosurfactants, Mannosylerythritol lipids." Journal of Oleo Science 57(2): 123–31.

147. Morita T, Konishi M, FukuokaT, Imura T, Kitamoto H and Kitamoto D. 2007b. "Characterization of the genus Pseudozyma by the formation of glycolipid biosurfactants, Mannosylerythritol lipids." FEMS Yeast Research 7(2): 286–92.

148. Morita T, Tokuma F, Tomohiro I and Dai K. 2009a. "Production of glycolipid biosurfactants by Basidiomycetous yeasts." Biotechnology and applied biochemistry 53: 39–49.

149. Morita T, Tokuma F, Tomohiro I, Dai K. 2013b. "Production of Mannosylerythritol lipids and their application in cosmetics." Applied Microbiology and Biotechnology 97(11): 4691–4700.

150. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. 2008. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". Nature Methods, 5(7), 621–628.

151. Mosher, J. J., Bowman, B., Bernberg, E. L., Shevchenko, O., Kan, J., Korlach, J., and Kaplan, L. a. 2014. "Improved performance of the PacBio SMRT technology for 16S rDNA sequencing". Journal of Microbiological Methods, 104, 59–60.

152. Mudge J. and Harrow J. 2016. "The State of play in higher eukaryote gene annotation." Nature Reviews Genetics 17(12): 758–72.

153. Mueller AN, Ziemann S, Treitschke S, Aßmann D, Doehlemann G. 2013. "Compatibility in the Ustilago maydis-maize interaction requires inhibition of host cysteine proteases by the fungal effector Pit2". PLoS Pathog. 9, e1003177.

154. Mueller G. and Schmit J. 2007. "Fungal Biodiversity: What Do We Know? What Can We Predict?" Biodiversity and Conservation 16(1): 1–5.

155. Müller,W.H., Bovenberg, R.A.L., Groothuis, M.H., Kattevilder, F., Smaal, E.B., van derVoort, L.H., andVerkleij, A.J. 1992. "Involvement of microbodies in penicillin biosynthesis". Biochim Biophys Acta 1116: 210–213

156. Müller,W.H., der Krift, T.P., Krouwer, A.J., Wösten, H.A., van der Voort, L.H., Smaal, E.B., and Verkleij, A.J. 1991. "Localization of the pathway of the penicillin biosynthesis in Penicillium chrysogenum". EMBO J 3: 489–495.

157. Murzin A. G., Brenner S. E., Hubbard T., Chothia C. 1995. "SCOP: a structural classification of proteins database for the investigation of sequences and structures". J. Mol. Biol. 247, 536-540.

158. National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988].
159. Ness SA.1999. "Myb binding proteins: regulators and cohorts in transformation". Oncogene 18:3039–3046.
160. Niu Y, Fan L, Gu D1, Wu J1, Chen Q. 2017. "Characterization, enhancement and modelling of Mannosylerythritol lipid production by fungal endophyte Ceriporia lacerate CHZJU." Food Chemistry 228: 610–17.
161. Nugent K, Choffe K and Saville B. 2004. "Gene expression during Ustilago maydis diploid filamentous growth: EST library creation and analyses." Fungal Genetics and Biology 41: 349–60 of mannosylerythritol lipid-B. Carbohydr Res.
162. Nyrén P, Lundin A. 1985. "Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis". Anal. Biochem. 509 (1985) 504–509
163. Okmen B, Collemare J, Griffiths S, van der Burgt A, Cox R, de Wit PJ. 2014. "Functional analysis of the conserved transcriptional regulator CfWor1 in Cladosporium fulvum reveals diverse roles in the virulence of plant pathogenic fungi". Molecular microbiology. 92(1):10–27.
164. Paraszkiewicz K., Dlugoński J.2003. "Microbial biosurfactants synthesis and applications" (Pol). Biotechnologia, 4, 82- 91, 2003.
165. Paterson RRM. 2005. "Fungus or bacterium and vice versa?". Microbiology 151:641.
166. Petersen JH .2013. "The Kingdom of Fungi" Princeton Univ Press, Princeton, NJ.
167. Pfaffl, M. W. 2001. "A New mathematical model for relative quantification in real-time RT-PCR." Nucleic Acids Research 29(9): 45e–45.
168. Platt A, Langdon T, Arst HN Jr, Kirk D, Tollervey D, Sanchez JM, Caddick M. 1996. "Nitrogen metabolite signalling involves the C-terminus and the GATA domain of the Aspergillus transcription factor AREA and the 3' untranslated region of its mRNA." The EMBO journal 15(11): 2791–2801
169. Ponting, C.P., Schultz, J.,Milpetz, F. and Bork, P. 1995. "SMART: identification and annotation of domains from signalling and extracellular protein sequences". Nucleic Acids Res 1999; 27: 229-232.
170. Raistrick, H. 1950. "A region of biosynthesis". Proc. R. Soc. Lond. B Biol. Sci. 136, 481–
171. Rau U, Nguyen LA, Roeper H, Koch H, Lang S. 2005. "Fed-batch bioreactor production of Mannosylerythritol lipids secreted by Pseudozyma aphidis." Applied Microbiology and Biotechnology 68(5): 607–13.
172. Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Birol, I. 2010. "De novo assembly and analysis of RNA-seq data". Nature Methods, 7(11), 909–12.
173. Rodrigues L, Banat IM, Teixeira J and Oliveira R. 2006. "Biosurfactants: potential applications in medicine". J Antimicrob Chemother 57:609–618.
174. Rohan L, Neil S, Bleackley M, Dolan S, Shafee T. 2017. "Transcriptomics Technologies." PLoS Computational Biology 13(5): 1–23.
175. Saika A. 2014. "Draft genome sequence of the yeast Pseudozyma antarctica type strain JCM10317 , a Producer of the Glycolipid Biosurfactants ,." 2(5): 4–5.

176. Saika A., Koike, H., Fukuoka, T., Yamamoto, S., Kishimoto, T., and Morita, T. 2016. A gene cluster for biosynthesis of mannosylerythritol lipids consisted of 4-O--D-Mannopyranosyl-(2R,3S)-Erythritol as the sugar moiety in a basidiomycetous yeast Pseudozyma tsukubaensis. PLoS ONE, 11(6), 1–16.
177. Santhanam P, Thomma BP. 2013. "Verticillium dahliae Sge1 differentially regulates expression of candidate effector genes". Molecular plant-microbe interactions: MPMI. 2013; 26(2):249–56.
178. Sari M. Kanti A, Artika and Kusharyoto W. 2013. "Identification of Pseudozyma hubeiensis Y10BS025 as a potent producer of glycolipid biosurfactant Mannosylerythritol lipid". American Journal of Biochemistry and Biotechnology 9 (4): 430-437, 2013
179. Schadt E, S. Turner, A. Kasarskis. "A window into third-generation sequencing". Hum. Mol. Genet. 19 (2010) R227–R240.
180. Schirawski J, Mannhaupt G, Münch K, Brefort T, Schipper K, Doehlemann G and Kahmann R. 2008. "Pathogenicity Determinants in Smut". Science. 10–13.
181. Schultz H. 1991. Beta oxidation of fatty acids. Biochim Biophys Acta. Jan 28;1081(2):109-20.
182. Schuster M, Schweizer G and Kahmann R. 2016. "Comparative analyses of secreted proteins in plant pathogenic smut fungi and related basidiomycetes." Fungal Genetics and Biology.
183. Sharma R, Xia X, Riess K, Bauer R, Thines M.2015. "Comparative genomics including the early-diverging smut fungus Ceraceosorus bombacis reveals signatures of parallel evolution within plant and animal pathogens of fungi and Oomycetes." Genome Biology and Evolution 7(9): 2781–98.
184. Sigrist C, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A and Buche P. 2002. "PROSITE: A documented database using patterns and profiles as motif descriptors". Brief. Bioinfo. Nucl. Acids Res. VOL 3. N3. 265–274.
185. Smith D, Garcia-Pedrajas M, Gold S and Perlin M. 2003. "Isolation and characterization from pathogenic fungi of genes encoding ammonium permeases and their roles in dimorphism." Molecular Microbiology 50(1): 259–75.
186. Soberón-Chávez, G., Marier R. 2010. "Biosurfactants and Overview In SoberónChávez, G (Ed.). Biosurfactants: from genes to applications. (1-13) Berlin: Springer.Hewald, S.,
187. Spröte P, Brakhage A, and Hynes J. 2009. "Contribution of peroxisomes to penicillin biosynthesis in Aspergillus nidulans". Eukaryot Cell 8: 421–423.
188. Stanke M and Morgenstern B. 2005. "AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints". Nucleic Acids Research, 33(SUPPL. 2), 465–467.
189. Tanaka S, Brefort T, Neidig N, Djamei A, Kahnt J, Vermerris W, Koenig S, Feussner K, Feussner I, Kahmann R. 2014. A secreted Ustilago maydis effector promotes virulence by targeting anthocyanin biosynthesis in maize. ELife 3, e01355.
190. Taniguti LM, Schaker PD, Benevenuto J, Peters LP, Carvalho G, Palhares A, Quecine MC, Nunes FR, Kmit MC, Wai A, Hausner G, Aitken KS, Berkman PJ, Fraser JA, Moolhuijzen PM, Coutinho LL, Creste S, Vieira

ML, Kitajima JP and Monteiro-Vitorello CB. 2015. “Complete genome sequence of Sporisorium scitamineum and biotrophic interaction transcriptome with sugarcane.” PLoS ONE 10(6): 1–31.

191. Teichmann B, Linne U, Hewald S, Marahiel M and Bölker M. 2007. “A biosynthetic gene cluster for a secreted cellobiose lipid with antifungal activity from Ustilago maydis”. Molecular Microbiology, 66(September), 525–533.

192. Thompso Jr and Nozawa Y. 1972. “Lipids of protozoa: phospho- lipids and neutral lipids”. Annu Rev Microbiol 26:249–278.

193. Thompson JD, Higgins DG, Gibson TJ. 1994. “CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting. Position-specific gap penalties and weight matrix choice”. Nucleic Acids Res 22:4673–4680.

194. Thudi M, Li Y.P, Jackson S.A, May G.D and Varshney R.K. 2012. “Current state-of-art of sequencing technologies for plant genomics research”. Brief. Funct. Genomics 11, 3–11.

195. Timothy L. Bailey and Charles Elkan. 1994. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California,

196. Todd R.B, Fraser J.A, Wong K.H, Davis M.A. andHynes, M.J., 2005. “Nuclear accumulation of the GATA factor AreA in response to complete nitrogen starvation by regulation of nuclear export”. Eukaryot. Cell 4, 1646–1653.

197. Tollot M, Assmann D, Becker C, Altmüller J, Dutheil JY, Wegner CE, Kahmann R. 2016. “The WOPR protein ros1 is a master regulator of sporogenesis and late effector gene expression in the maize pathogen Ustilago maydis.” PLoS Pathogens 12(6): 1–37.

198. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley D and Pachter, L. 2012. “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”. Nature Protocols, 7(3), 562–78.

199. Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, van Baren M. J and Pachter, L. 2010. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. Nature Biotechnology, 28(5), 511–515.

200. Troy D, Remington J and Beringer P. 2006. *Remington: The Science and Practice of Pharmacy* (21st ed.). Philadelphia: Lippincott Williams & Wilkins. pp. 325–336, 886–87. ISBN 0-7817-4673-6.

201. Tunlid A, Talbot NJ. 2002. “Genomics of parasitic and symbiotic fungi”. Genomics.5:513–9.

202. Vogel HJ. 1964. ”Distribution of lysine pathways among fungi: evolutionary implications”. Am Nat 98:435–446.

203. Walton, J.D. 2000. “Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: an hypothesis”. Fungal Genet Biol 30: 167–171.

204. Wang L, Feng Z, Wang X, Wang X andZhang X. 2009. “DEGseq: An R Package for Identifying Differentially Expressed Genes from RNA-Seq Data.” Bioinformatics 26(1): 136–38.

205. Wang L, Wang S and Li W. 2012. "RSeQC: Quality control of RNA-Seq experiments." Bioinformatics 28(16): 2184–85.
206. Wang Q. M, Begerow D, Groenewald M, Liu X, Theelen B, Bai F. Y and Boekhout, T. 2015. "Multigene phylogeny and taxonomic revision of yeasts and related fungi in the Ustilaginomycotina". Studies in Mycology, 81, 55–83.
207. Weaver CT1, Hatton RD, Mangan PR, Harrington LE. 2007. "IL-17 family cytokines and the expanding diversity of effector T cell lineages". Annu. Rev. Immunol. 25: 821-852.
208. Wickham H. 2016. "ggplot2: Elegant Graphics for Data Analysis". Springer-Verlag New York.
209. Włodzimierz B, Adamczak M, Tomasik J and Płaszczyk M. 2004. "Application of Oil Refinery Waste in the Biosynthesis of Glycolipids by Yeast. Bioresource Technology". 95(1): 15–18.
210. Wollenberg T and Schirawski, J. 2014. "Comparative genomics of plant fungal pathogens: the Ustilago-Sporisorium paradigm". PLoS Pathogens, 10(7), 1–3.
211. Wray S, Nimtz M and S. Lang. 1999. "Glycolipids of the smut fungus Ustilago maydis from cultivation on renewable resources." Applied Microbiology and Biotechnology 51(1): 33–39.
212. Xu Xihui, Qin He, Chen Chen, and Chulong Zhang. 2016. "Differential Communications between fungi and host plants revealed by secretome analysis of phylogenetically related endophytic and pathogenic fungi." PLoS ONE 11(9): 1–16.
213. Yanagishita H, Kenji H and Kitamoto H. 1998. "Contribution of a chain-shortening pathway to the biosynthesis of the fatty acids of mannosylerythritol lipid (biosurfactant) in the yeast Candida antarctica : Effect of ω -Oxidation Inhibitors on Biosurfactant Synthesis." 20(9): 813–18.
214. Yandell, M. and Ence, D. 2012. "A beginner's guide to eukaryotic genome annotation". Nature Reviews. Genetics, 13(5), 329–42.
215. Yoshida S, Morita T, Shinozaki Y, Watanabe T, Sameshima-Yamashita Y, Koitabashi, M and Kitamoto H. 2014. "Mannosylerythritol lipids secreted by phyllosphere yeast Pseudozyma antarctica is associated with its filamentous growth and propagation on plant surfaces". Applied Microbiology and Biotechnology, 98(14), 6419–6429.
216. Zavala-Moreno A, Arreguin-Espinosa R, Pardo J, Romero-Aguilar L, Guerra-Sánchez G . 2014. "Nitrogen source affects glycolipid production and lipid accumulation in the phytopathogen fungus Ustilago maydis. Advances in Microbiology 04(13): 934–44.

## LIST OF APPENDIX

Appendix. 2-1. TopHat script used

```
#Script to run in bash all libraries from both fermentations

for f in 3_1.1 3_2.1 3_3.1 4_1.1 4_2.1 4_3.1 3_1.2 3_2.2 3_3.2 3_4.2 3_5.2
4_1.2 4_2.2 4_3.2 4_4.2 4_5.2 ; do

      tophat2 \
            --num-threads 40 \
            -o F${f}cufflinks \
            --min-intron-length 20 \
            --max-intron-length 5000 \
            --library-type fr-firststrand \
            --mate-std-dev 33 \
            --mate-inner-dist 102 \
            /pub39/ext/stefany/RNA_Seq/PGRAM \
            /pub39/ext/stefany/Trimmed/F${f}/*_R1_001.fastq
            /pub39/ext/stefany/Trimmed/F${f}/*_R2_001.fastq

      done;
done;
```

Appendix. 2-2. Braker default parameters

```
#script bed_to_gff default parameters
bet_to_gff.pl  --bed name_file.junctions.bed --seq fasta_file --v --gff output_file

#Braker example code
perl braker.pl --species=file_name --cores=25 --AUGUSTUS_CONFIG_PATH=/user/path --
BAMTOOLS_PATH=/user/path/ --GENEMARK_PATH=/user/path/ --genome= fasta_file --
-hints=hints_file --workingdir=/user/directory/ --skipGeneMark-ET
```

Appendix. 2-3.Hydrolase family genes in *P. graminicola*

| GENE | ASSOCIATED FUNCTION |
|---|---|
| g241 | Dolichyl-diphospho oligosaccharide glycosyltransferase subunit stt3 |
| g1476 | Glycosyl hydrolase five-bladed beta-propellor domain |
| g2622 | Uracil-DNA glycosylase Short |
| g3137 | Uncharacterized glycosyl hydrolase YIR007W |
| g3264 | DNA-3-methyladenine glycosylase 1 |
| g3713 | Glycosyl hydrolase, five-bladed beta-propellor domain |
| g5057 | Uncharacterized glycosyl hydrolase YIR007W |
| g5101 | Glycosyl hydrolase, five-bladed beta-propellor domain |
| g5188 | Adenine DNA glycosylase |

| | |
|---|---|
| g5448 | Dolichyl-diphosphooligosaccharide- glycosyltransferase subunit 3 |
| g5521 | N-glycosylase DNA lyase |
| g6077 | mismatch-specific uracil DNA glycosylase |
| g6339 | Dolichyl-diphospho oligosaccharide-- glycosyltransferase subunit wbp1 |
| g6470 | Probable_dolichyl-diphospho oligosaccharide-- glycosyltransferase subunit |
| g6584 | Dolichyl-diphosphooligosaccharide-- glycosyltransferase subunit 1 |

Appendix. 2-4. Oxidoreductase group family in *P. graminicola*

| GENE | ASSOCIATED FUNCTION |
|---|---|
| g1082 | Disulfide-isomerase erp38 Short |
| g1390 | 37S ribosomal mitochondrial |
| g2604 | Phytanoyl-CoA dioxygenase |
| g2755 | NAD-dependent |
| g3083 | Ergothioneine biosynthesis 1 |
| g3976 | Glutaredoxin-1 |
| g4079 | Flavo YCP4 |
| g4191 | Thioredoxin Short |
| g4927 | Probable cytosine deaminase |
| g5166 | NADH dehydrogenase (ubiquinone) 1 alpha subcomplex subunit 5 |
| g5765 | Thioredoxin- mitochondrial Short |
| g5943 | Thioredoxin 1 |
| g6424 | NADPH-dependent methylglyoxal reductase GRE2 |

Appendix. 2-5.*P. graminicola* orthologs to grass effector clusters from related smut fungi. Table displaying genes comprised in each cluster.

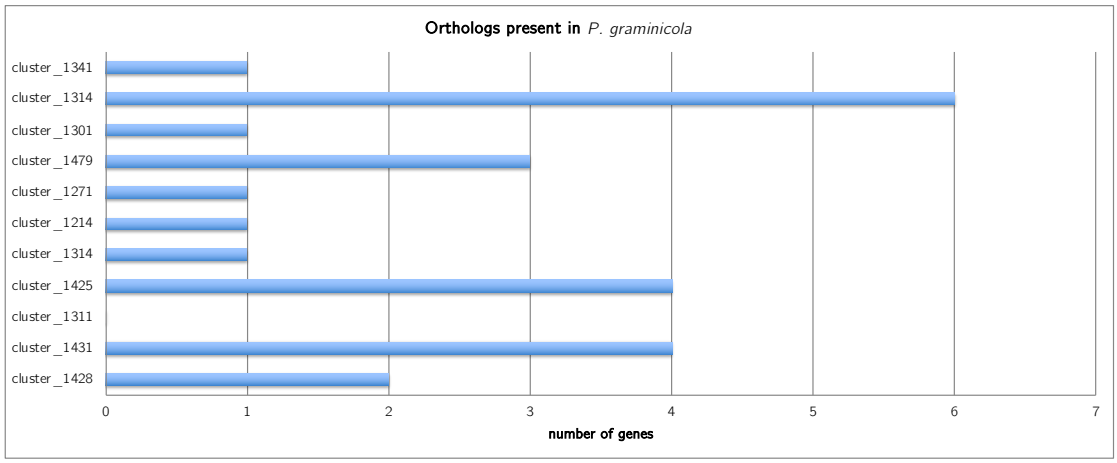



Table with blastp results from *P. graminicola* proteins to effectors specific to smut fungi infecting grass at a cut-off value of $10^{-5}$. Clusters taken from Schuster et al. 2016.

| Cluster | Protein_ID | PGRAM_ID | %id | Alignment length |
|---|---|---|---|---|
| 1250 | UMAG10972 | g928.t1 | 34.75 | 541 |
| 1250 | SPSC03771 | g928.t1 | 42.07 | 523 |
| 1250 | sr16775 | g928.t1 | 42.4 | 533 |
| 1250 | UHOR08856 | g928.t1 | 28.8 | 573 |
| 1261 | SPSC01549 | g412.t1 | 35.78 | 109 |
| 1261 | sr16441 | g412.t1 | 36.54 | 104 |
| 1267 | UMAG12197 | g5605.t1 | 63.3 | 109 |
| 1267 | SPSC06609 | g5605.t1 | 66.67 | 129 |
| 1267 | sr13927 | g5605.t1 | 69.16 | 107 |
| 1267 | UHOR04531 | g5605.t1 | 61.79 | 123 |
| 1270 | UMAG04039 | g5272.t1 | 43.59 | 117 |
| 1270 | SPSC05323 | g5272.t1 | 48.75 | 160 |
| 1270 | sr14946 | g5272.t1 | 51.16 | 129 |
| 1270 | UHOR13428 | g5272.t1 | 40.24 | 164 |
| 1280 | UMAG00104 | g2784.t1 | 47.46 | 295 |
| 1280 | UMAG00104 | g2784.t1 | 41.43 | 70 |
| 1280 | UMAG00104 | g4880.t1 | 29.31 | 174 |
| 1280 | SPSC02335 | g2784.t1 | 52.11 | 641 |
| 1280 | SPSC02335 | g4880.t1 | 29.38 | 160 |
| 1280 | sr11444 | g2784.t1 | 60.79 | 454 |
| 1280 | sr11444 | g2784.t1 | 66.2 | 71 |
| 1280 | sr11444 | g4880.t1 | 29.27 | 164 |
| 1280 | UHOR00167 | g2784.t1 | 42 | 300 |
| 1280 | UHOR00167 | g4880.t1 | 30.49 | 164 |
| 1298 | UMAG03822 | g5050.t1 | 79.13 | 115 |
| 1298 | SPSC05184 | g5050.t1 | 81.9 | 116 |
| 1298 | sr14724 | g5050.t1 | 82.76 | 116 |
| 1298 | UHOR05809 | g5050.t1 | 68.1 | 116 |
| 1306 | UMAG12127 | g4337.t1 | 66.38 | 116 |
| 1306 | SPSC04538 | g4337.t1 | 73.11 | 119 |
| 1306 | sr12428 | g4337.t1 | 64.1 | 117 |
| 1306 | UHOR01705 | g4337.t1 | 57.26 | 117 |
| 1307 | UMAG10274 | g1204.t1 | 85.95 | 121 |
| 1307 | SPSC03183 | g1204.t1 | 91.74 | 121 |
| 1307 | sr10650 | g1204.t1 | 93.39 | 121 |
| 1307 | UHOR02844 | g1204.t1 | 81.82 | 121 |
| 1341 | sr16558 | g2300.t1 | 26.85 | 149 |
| 1341 | sr16560 | g2300.t1 | 40.3 | 67 |
| 1357 | UMAG02298 | g570.t1 | 27.99 | 268 |
| 1357 | SPSC01709 | g569.t1 | 35.51 | 321 |
| 1357 | SPSC01710 | g570.t1 | 36.81 | 182 |

| | | | | |
|---|---|---|---|---|
| 1357 | sr13495 | g570.t1 | 40.97 | 144 |
| 1376 | UMAG03202 | g4122.t1 | 42.2 | 109 |
| 1376 | UMAG10403 | g4122.t1 | 49.09 | 110 |
| 1376 | SPSC04237 | g4122.t1 | 33.98 | 103 |
| 1376 | SPSC04238 | g4122.t1 | 63.24 | 136 |
| 1376 | sr11101 | g4122.t1 | 33.65 | 104 |
| 1376 | sr11102 | g4122.t1 | 70.43 | 115 |
| 1376 | UHOR04986 | g4122.t1 | 36.89 | 122 |
| 1390 | UMAG00792 | g3684.t1 | 47.77 | 157 |
| 1390 | UMAG00793 | g3684.t1 | 23.4 | 141 |
| 1390 | SPSC02060 | g3684.t1 | 56.69 | 157 |
| 1390 | sr12084 | g3684.t1 | 54.29 | 175 |
| 1390 | sr12085 | g3684.t1 | 45.45 | 176 |
| 1390 | UHOR01209 | g3684.t1 | 26.97 | 178 |
| 1409 | UMAG04035 | g5271.t1 | 30.43 | 115 |
| 1409 | UMAG11058 | g5271.t1 | 27.08 | 96 |
| 1409 | SPSC05325 | g5271.t1 | 25 | 108 |
| 1409 | SPSC05326 | g5271.t1 | 43.2 | 125 |
| 1425 | UMAG00558 | g5523.t1 | 30.28 | 109 |
| 1425 | UMAG00558 | g5523.t1 | 21.66 | 157 |
| 1425 | UMAG01300 | g5802.t1 | 27.07 | 133 |
| 1425 | UMAG01302 | g5523.t1 | 34.23 | 111 |
| 1425 | UMAG01302 | g5523.t1 | 26.42 | 159 |
| 1425 | SPSC01738 | g5523.t1 | 39.81 | 108 |
| 1425 | SPSC01738 | g5523.t1 | 24.39 | 164 |
| 1425 | SPSC01739 | g5523.t1 | 31.4 | 121 |
| 1425 | SPSC02294 | g5523.t1 | 35.96 | 89 |
| 1425 | SPSC02294 | g5523.t1 | 23.98 | 171 |
| 1425 | SPSC03604 | g5523.t1 | 45.88 | 170 |
| 1425 | SPSC03604 | g5523.t1 | 43.51 | 154 |
| 1425 | sr11400 | g5523.t1 | 29.61 | 152 |
| 1425 | sr13522 | g5523.t1 | 38.46 | 91 |
| 1425 | sr13524 | g5523.t1 | 35.29 | 85 |
| 1425 | sr13524 | g5523.t1 | 33.7 | 92 |
| 1425 | sr13525 | g5523.t1 | 33.33 | 75 |
| 1425 | sr13525 | g5523.t1 | 37.31 | 67 |
| 1425 | sr20001 | g5523.t1 | 38.81 | 134 |
| 1425 | sr20001 | g5523.t1 | 30.63 | 160 |
| 1425 | UHOR01947 | g5524.t1 | 27.97 | 118 |
| 1428 | SPSC00075 | g251.t1 | 26.19 | 168 |
| 1428 | SPSC00077 | g251.t1 | 29.47 | 190 |
| 1428 | SPSC00078 | g252.t1 | 54.27 | 199 |
| 1428 | SPSC00079 | g253.t1 | 50.25 | 199 |

| | | | | |
|---|---|---|---|---|
| **1428** | sr10050 | g251.t1 | 31 | 200 |
| **1428** | sr10052.2 | g253.t1 | 61.9 | 189 |
| **1431** | SPSC00094 | g259.t1 | 52.05 | 171 |
| **1431** | SPSC00097 | g260.t1 | 32.39 | 142 |
| **1431** | SPSC00097 | g261.t1 | 36.45 | 107 |
| **1431** | sr10073 | g262.t1 | 61.4 | 171 |
| **1431** | sr10075 | g260.t1 | 29.17 | 144 |
| **1431** | sr10075 | g261.t1 | 33.33 | 144 |
| **1431** | sr10079 | g260.t1 | 32.64 | 144 |
| **1431** | sr10079 | g261.t1 | 30.25 | 119 |
| **1431** | sr20014 | g259.t1 | 60.57 | 175 |
| **1431** | UHOR08134 | g259.t1 | 40.88 | 137 |
| **1452** | UMAG05930 | g2299.t1 | 38.52 | 135 |
| **1452** | SPSC06087 | g2299.t1 | 40.87 | 115 |
| **1452** | sr13344 | g2299.t1 | 32.81 | 64 |
| **1452** | sr16553 | g2299.t1 | 30.94 | 139 |
| **1452** | UHOR03426 | g2299.t1 | 29.75 | 158 |
| **1464** | UMAG03223 | g4144.t1 | 41.22 | 148 |
| **1464** | UMAG03223 | g4145.t1 | 58.97 | 117 |
| **1464** | UMAG03223 | g4143.t1 | 26.67 | 150 |
| **1464** | UMAG03223 | g4146.t1 | 27.56 | 156 |
| **1464** | UMAG12216 | g4146.t1 | 32.89 | 152 |
| **1464** | UMAG12216 | g4144.t1 | 28.86 | 149 |
| **1464** | SPSC04259 | g4143.t1 | 34.52 | 168 |
| **1464** | SPSC04260 | g4143.t1 | 40.22 | 179 |
| **1464** | SPSC04261 | g4143.t1 | 38.64 | 132 |
| **1464** | SPSC04263 | g4143.t1 | 48.98 | 49 |
| **1464** | SPSC04264 | g4143.t1 | 41.98 | 131 |
| **1464** | SPSC04265 | g4143.t1 | 36.91 | 149 |
| **1464** | SPSC04266 | g4143.t1 | 48.12 | 160 |
| **1464** | SPSC04267 | g4143.t1 | 41.61 | 161 |
| **1464** | SPSC04268 | g4143.t1 | 40 | 155 |
| **1464** | SPSC04270 | g4144.t1 | 64.67 | 184 |
| **1464** | SPSC04270 | g4145.t1 | 42.62 | 122 |
| **1464** | SPSC04270 | g4146.t1 | 28.78 | 139 |
| **1464** | SPSC04270 | g4143.t1 | 31.43 | 140 |
| **1464** | SPSC04271 | g4145.t1 | 55.83 | 120 |
| **1464** | SPSC04271 | g4144.t1 | 37.09 | 151 |
| **1464** | SPSC04271 | g4146.t1 | 28.46 | 130 |
| **1464** | SPSC04271 | g4143.t1 | 26.14 | 176 |
| **1464** | SPSC04272 | g4146.t1 | 64.95 | 214 |
| **1464** | SPSC04272 | g4144.t1 | 30.61 | 147 |
| **1464** | SPSC04273 | g4144.t1 | 32.21 | 149 |

| | | | | |
|---|---|---|---|---|
| **1464** | SPSC04273 | g4146.t1 | 35.1 | 151 |
| **1464** | sr02614 | g4143.t1 | 45.45 | 187 |
| **1464** | sr11130 | g4143.t1 | 33.62 | 116 |
| **1464** | sr11132 | g4143.t1 | 41.77 | 158 |
| **1464** | sr11133 | g4143.t1 | 38.89 | 162 |
| **1464** | sr14220 | g4145.t1 | 58.2 | 122 |
| **1464** | sr14220 | g4144.t1 | 39.35 | 155 |
| **1464** | sr14220 | g4143.t1 | 28.3 | 159 |
| **1464** | sr14220 | g4146.t1 | 30.71 | 127 |
| **1464** | sr14222 | g4144.t1 | 35.44 | 158 |
| **1464** | sr14222 | g4146.t1 | 32.89 | 152 |
| **1464** | UHOR06702 | g4144.t1 | 46.62 | 148 |
| **1464** | UHOR06702 | g4146.t1 | 27.74 | 137 |
| **1464** | UHOR06702 | g4145.t1 | 34.15 | 123 |
| **1464** | UHOR08826 | g4144.t1 | 43.36 | 143 |
| **1464** | UHOR06051 | g4144.t1 | 42.36 | 144 |
| **1464** | UHOR06051 | g4145.t1 | 32.5 | 120 |
| **1464** | UHOR06051 | g4146.t1 | 27.21 | 147 |
| **1464** | UHOR08252 | g4144.t1 | 46.31 | 149 |
| **1464** | UHOR08252 | g4146.t1 | 34.18 | 158 |
| **1464** | UHOR08252 | g4145.t1 | 35.29 | 119 |
| **1464** | UHOR06234 | g4144.t1 | 43.62 | 149 |
| **1464** | UHOR06234 | g4146.t1 | 27.54 | 138 |
| **1464** | UHOR06234 | g4145.t1 | 36.67 | 120 |
| **1464** | UHOR06803 | g4143.t1 | 41.05 | 95 |
| **1464** | UHOR04923 | g4144.t1 | 31.28 | 195 |
| **1464** | UHOR04923 | g4146.t1 | 32.89 | 149 |
| **1464** | UHOR04923 | g4145.t1 | 41.75 | 103 |
| **1464** | UHOR04922 | g4144.t1 | 31.18 | 170 |
| **1464** | UHOR04922 | g4146.t1 | 29.65 | 172 |
| **1464** | UHOR04990 | g4143.t1 | 30.67 | 150 |
| **1464** | UHOR15214 | g4146.t1 | 37.76 | 196 |
| **1464** | UHOR15214 | g4144.t1 | 29.05 | 148 |
| **1464** | UHOR04675 | g4144.t1 | 39.61 | 154 |
| **1464** | UHOR04675 | g4145.t1 | 44.54 | 119 |
| **1464** | UHOR04676 | g4144.t1 | 33.51 | 191 |
| **1464** | UHOR04676 | g4146.t1 | 36.27 | 102 |
| **1464** | UHOR04676 | g4145.t1 | 42.27 | 97 |
| **1464** | UHOR04736 | g4144.t1 | 37.76 | 143 |
| **1464** | UHOR04736 | g4146.t1 | 29.55 | 176 |
| **1464** | UHOR04736 | g4145.t1 | 39.83 | 118 |

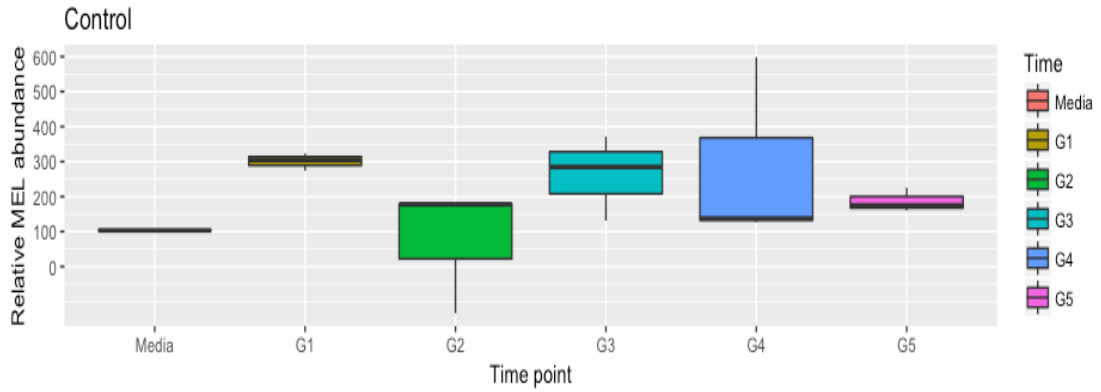
Appendix 3-1. Pattern file used for spectra designation of metabolites in *P. graminicola* fermantations.

```
PATTERN     = YEAST_MEDIA
GROUP       = IIB
DESCRIPTION = IN_HOUSE_FERMENTATIONS
AUTHOR      = SSG
DIM         = 2
ORIGIN      = 1
ITEMS       = 31
0.0000  0.0000  7.2719  7.2590  0  A_B (Chloroform)
0.0000  0.0000  5.5480  5.4590  0  C_D
0.0000  0.0000  5.4380  5.2840  0  FA_1
0.0000  0.0000  5.2780  5.1560  0  E_F
0.0000  0.0000  5.1530  5.0700  0  F_G
0.0000  0.0000  5.0850  4.9570  0  G_H
0.0000  0.0000  4.9550  4.8890  0  H_I
0.0000  0.0000  4.8800  4.8090  0  I_J
0.0000  0.0000  4.7680  4.7290  0  J_K
0.0000  0.0000  4.3640  4.2870  0  L_M_mannose
0.0000  0.0000  4.2690  4.1980  0  M_N_mannose
0.0000  0.0000  4.0790  3.9420  0  O_P_mannose
0.0000  0.0000  3.9180  3.7940  0  P_Q_mannose
0.0000  0.0000  3.7940  3.7230  0  Q_R_mannose
0.0000  0.0000  3.7100  3.6280  0  R_S_mannose
0.0000  0.0000  3.6280  3.4910  0  S_T_mannose
0.0000  0.0000  3.4910  3.4290  0  T_U_mannose
0.0000  0.0000  3.4290  3.3700  0  MEL-related
0.0000  0.0000  2.6900  2.6900  0  FA_2
0.0000  0.0000  2.4820  2.3870  0  X_Y
0.0000  0.0000  2.3870  2.3270  0  FA_3
0.0000  0.0000  2.3210  2.2680  0  Z_AA
0.0000  0.0000  2.2680  2.1790  0  AA_BB
0.0000  0.0000  2.1730  2.1140  0  BB_CC
0.0000  0.0000  2.1140  1.9350  0  FA_4
0.0000  0.0000  1.6860  1.5850  0  FA_5
0.0000  0.0000  1.5790  1.5020  0  EE_FF
0.0000  0.0000  1.5020  1.1100  0  FA_6
0.0000  0.0000  0.9983  0.9315  0  GG_HH
0.0000  0.0000  0.9190  0.8466  0  FA_7
0.0000  0.0000  0.8466  0.7513  0  II_JJ
```

Appendix 3-2. Washing cell methods used to extract internal metabolites from *P. graminicola*

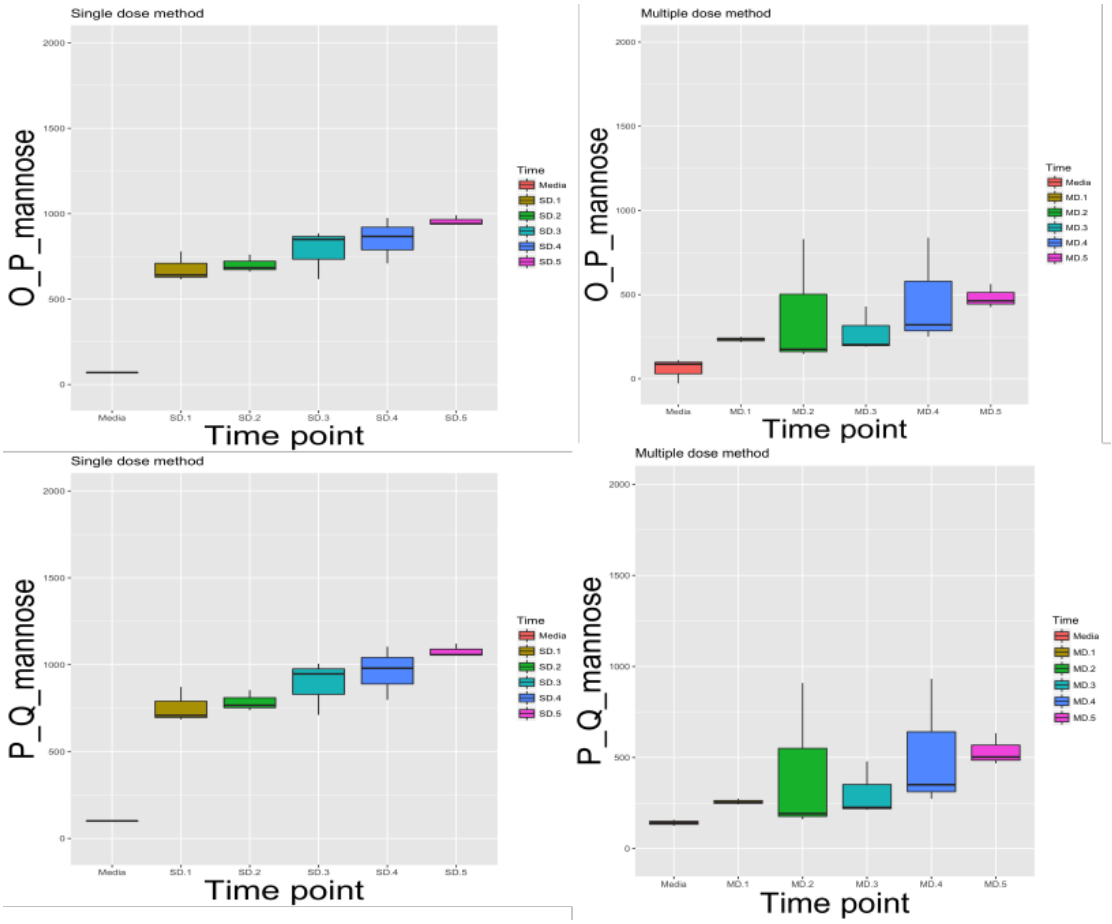
| Method | Wash | GENERAL WASH | Harvesting | Storage | Extraction reagent | Extraction technique |
|---|---|---|---|---|---|---|
| 1 | Triton X-100 (30 min) | PBS 4X | Filter | Liq N2 | Acetronile:Water (50:50) | Sonicate (30 sec on 30 sec off 3X) |
| 2 | Triton X-100 (30 min) | | Cenrifuge | -80C | ETOH 75% | Vortex 12X 30 sec on 30 sec off |
| 3 | Triton X-100 (30 min) | | Cenrifuge | Liq N2 | Acetronile:Water (50:50) | Sonicate (30 sec on 30 sec off 3X) |
| 4 | Triton X-100 (30 min) | | Cenrifuge | MeOH 200 uL + Liq N2 | Methanol 200 uL + Liq N2 | CHCl3:MeOH (Chloroform: Methanol) Liq N2 1 min, ice 2 min 5X |
| 5 | Triton X-100 (30 min) | | Cenrifuge | Liq N2 | Acetronile:Water (50:50) | Sonicate (30 sec on 30 sec off 15X) |
| 6 | PBS 1X | | Cenrifuge | Liq N2 | Acetronile:Water (50:50) | Sonicate (30 sec on 30 sec off 3X) |
| 7 | SDS 0.1% | | Cenrifuge | Liq N2 | Acetronile:Water (50:50) | Sonicate (30 sec on 30 sec off 3X) |
| 8 | Triton X-100 (30 min) | | Cenrifuge | Liq N2 | ETOH 75% | ETOH 75% + vortex 12X 30 sec on, 30 sec off |

Appendix 3-3. Relative abundance box plot for MEL-related metabolite flask fermentation for *P. graminicola* after 117 hours. N=3. Control flask (G) with no addition of FA. Numbers refer to time points: 1=24 h, 2= 48, 3= 72h, 4= 96, 5= 117h.




Appendix 3-4. Relative abundance box plots of normalised MEL and mannose-related metabolites over a 117 time course flask fermentation under two different feeding regimes.

MD= multiple dose regimen. SD= Single dose regimen. 1=24 h, 2=48 h, 3= 72 h, 4= 96 h, 5= 117h. N=3.

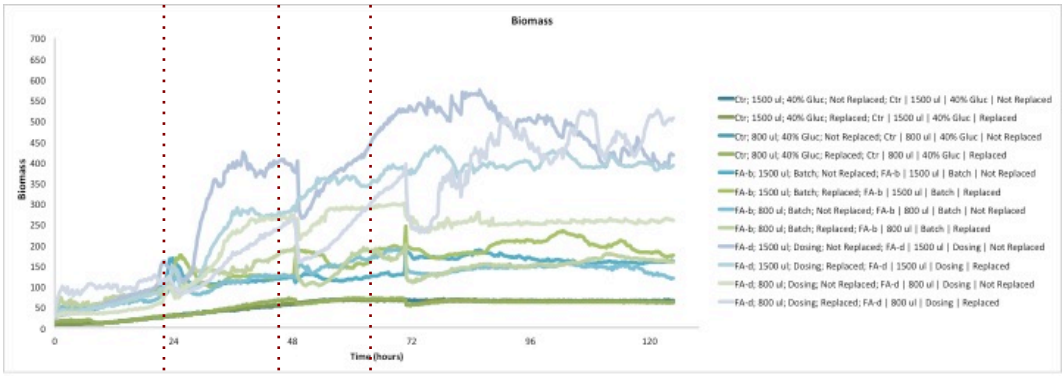


Appendix 3-5. Extended key for PCA growth rates.

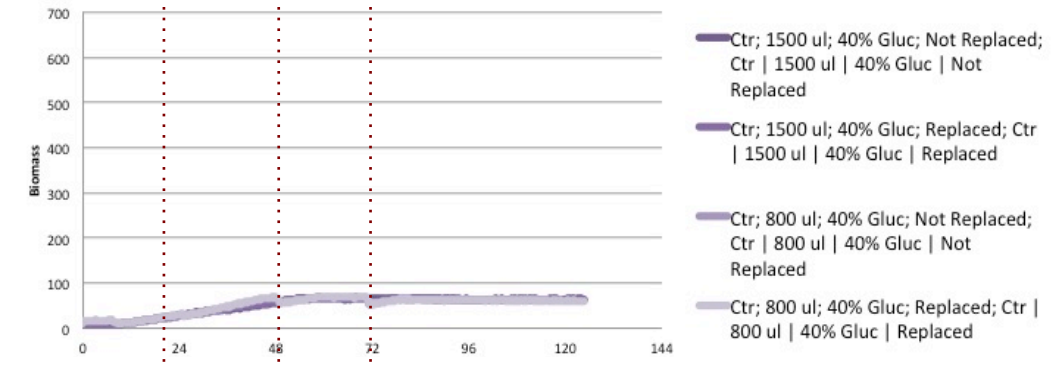
| Key | Method |
|---|---|
| A | Control 1500 ul glucose no replace |
| B | Control 1500 ul glucose replaced |
| C | Control 1500 ul glucose untouched |
| D | Control 800 ul glucose no replace |
| E | Control 800 ul glucose replaced |
| F | Control 800 ul glucose untouched |
| G | FA batch 1500 ul no replace |
| H | FA batch 1500 ul replaced |
| I | FA batch 1500 ul untouched |
| J | FA batch 800 ul no replace |
| K | FA batch 800 ul replaced |
| L | FA batch 800 ul untouched |
| M | FA shot 1500 ul no replace |
| N | FA shot 1500 ul replaced |
| O | FA shot 800 ul no replace |
| P | FA shot 800 ul replaced |

Appendix 3-6. Biomass records for 72 hours sampling (every 24 hours) plus 48 hours of recoding (no sampling). A= Joined data: controls and treatments, B= Control (no FA), C= Single dose regimen, D= Multiple dose regimen. Sampled points depicted with red dotted line.

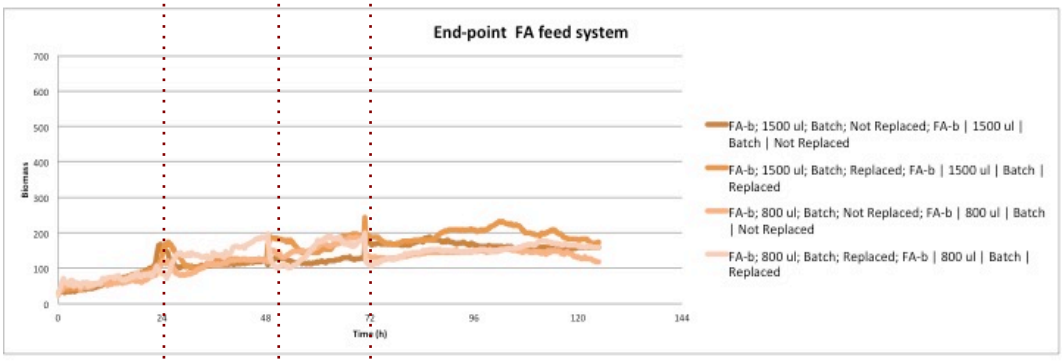
A



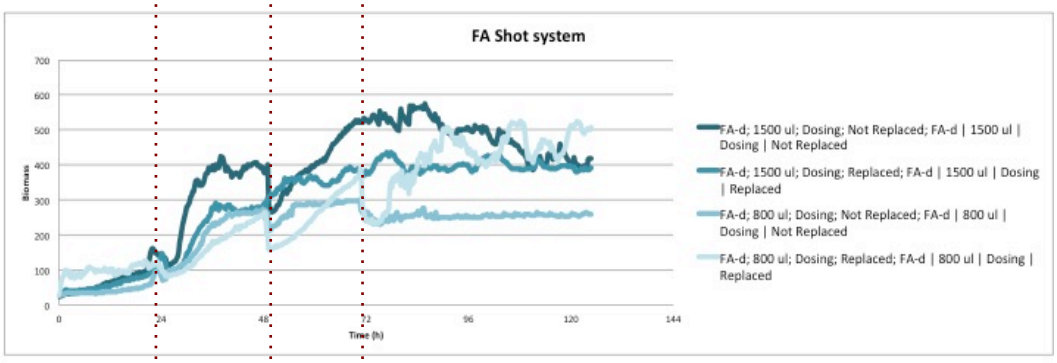

B




C




D



Appendix. 3-7. Standarised protocol to semi-quantify MELs from media samples on a FA background

Sample processing

1) Take between 200 to 600 microliters of media
2) Lyophilise samples over night
3) Add 600 microliters of deuterated chloroform to lyophilised
4) Vortex mix thoroughly1-2 min
5) Spin at 10000 g at 10 °C for 10. minutes
6) Transfer carefully volume, avoiding any undissolved particle to 5 mm diameter NMR tubes

Data analysis

7) Perform QC based on smoothness of chloroform peak at 7.26 ppp position
8) Acquire spectral data (area under the curve) using the pattern file and implement Amix to transform the spectral signal to concentration values, better known as bucket table, using manufacturer instructions and only spectra which passed QC.
9) Normalise spectral values exported to a csv format by Amix by using an R script developed by the CBF (University of Liverpool) to normalise peak intensity values (bucket table) to the chloroform peak(attached at the end of this protocol).
10) Utilise MetaboAnalyst, an online ensuite to scale and transform the normalised values by performing a pareto scaling and a cube root transformation.

R script

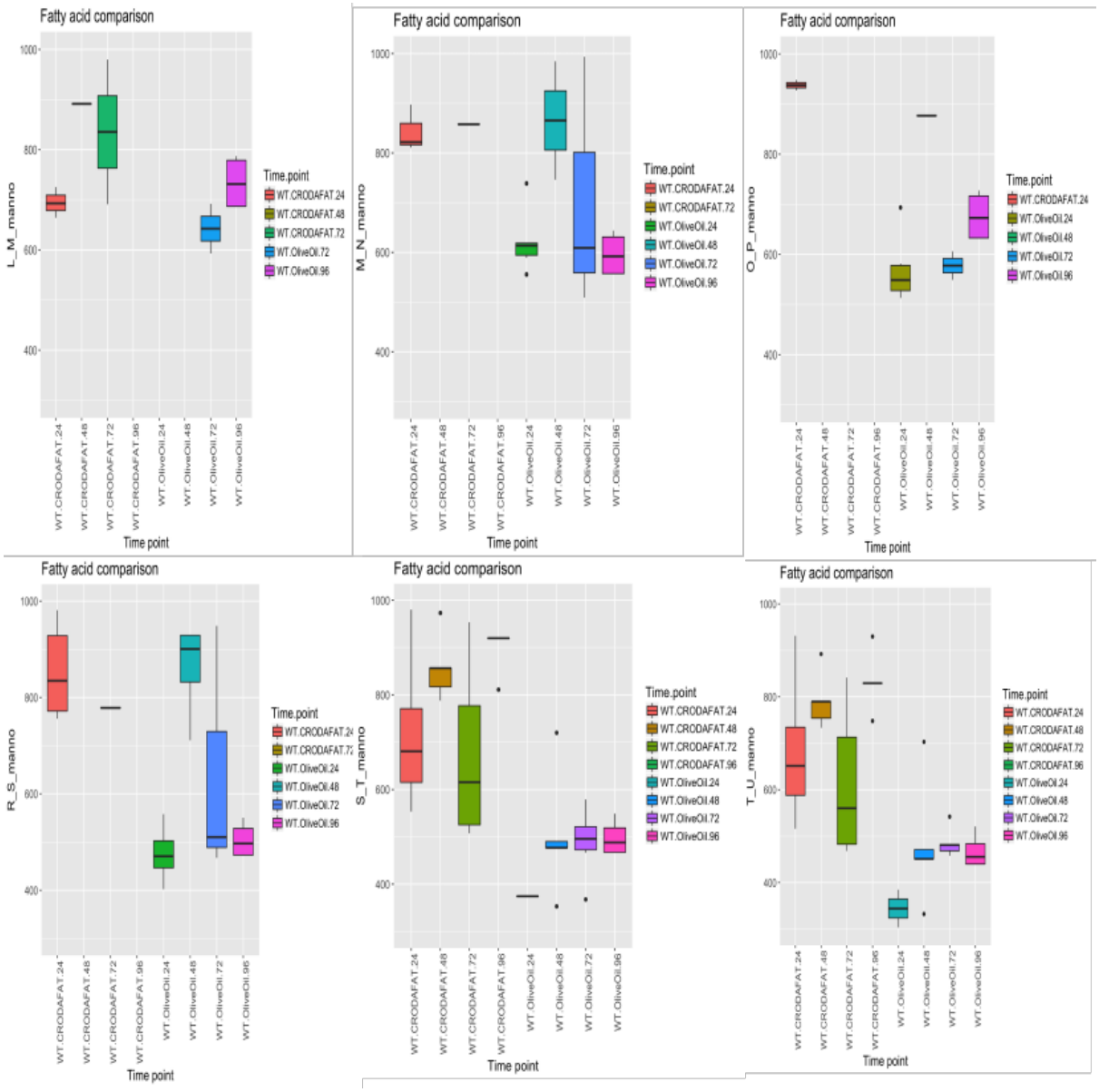
```
# Eva Caamano Gutierrez and Arturas Grauslys, 2017.
# This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
# https://creativecommons.org/licenses/by-nc-sa/4.0/

# R script for data normalisation and scaling
# Procedures are performed by calling the do_norm_scale() function with parameters:
#
# - data: a data frame with samples in the rows and variables (bins) in the columns.
#  The first column in the data (not counting sample names) should be a grouping variable
#
# - normalisation: a normalisation method to be used.
#     Available methods: "PQN" - Probabilistic quotient normalisation
#"TotArea" - normalisation by the total area under the curve
```

```
#"Bin" - normalisation by 1 bin in the dataset
#
# - bin: the number of column in the dataset by which normalisation has to be performed (only used for "Bin" normalisation)
#
# - scaling: a scaling method to be used.
#      Available methods: "Auto" - mean centering and scaling by the standard deviation
#                         "Pareto" - mean centering and scaling by the square root of the standard deviation
#                          "Range" - mean centering and scaling such that the data ranges from 0 to 1.
#                         "Mean" - mean centering

NMRMetab_norm_scale = function(data, normalisation = 'None', bin = NA, scaling = 'None'

  #separate data from groups
  data_ = as.matrix(data[,3:ncol(data)])
  grp = as.factor(data[,2])

  # apply normalisation
  if (normalisation == 'None'
    cat('Normalisation: None\n'
     else if (normalisation == 'PQN'
    data_ = PQN(data_)
    cat 'Normalisation: PQN\n'
  } else if (normalisation == 'TotArea'
    data_ = TotArea(da
         'Normalisation: Total area\n'
  } else if (normalisation == 'Bin'){
    if (!is.na(bin)){
      data_ = NormByBin(data_, bin)
         'Normalisation: Bin\n'
    cat(sprintf('Selected bin: %s \n'  as.character(bin)))
    } else {
      print(paste('Method does not exist: ', normalisation, sep=''))
    }
  }

  #apply scaling
  if (scaling == 'None'){
    cat('Scaling: None\n'
```

```
  } else if (scaling == 'Auto'
    data_ = apply(data_, 2, AutoScale)
    cat 'Scaling: Auto\n'
  } else if (scaling == 'Pareto'
    data_ = apply(data_, 2, ParetoScale)
    cat 'Scaling: Pareto\n'
  } else if (scaling == 'Range') {
    data_ = apply(data_, 2  RangeScale)
        'Scaling: Range\n'
  } else if (scaling == 'Mean') {
    data_ = scale(data_, center=T, scale=F)
    cat('Scaling: Mean\n'
  } else {
    cat(sprint 'Method does not exist: %s \n'
  }

  dataLabs = data[,1:2]
  out
  #data_ = cbind(1:nrow(data_), grp, data_)
  #out_data = as.data.frame(data_)

  return(out_data)
}

# scaling funcions
AutoScale<-function(x){
  (x - mean(x))/sd(x, na.rm=T)
}

ParetoScale<-function(x){
  (x - mean(x))/sqrt(sd(x, na.rm=T))
}

RangeScale<-function(x){
  if(max(x) == min(x)){
    x
  }else{
    (x - mean(x))/(max(x)-min(x))
  }
}

# normalisation
```

```
PQN <- function(data, loc = "median"

  if (loc == "mean"
    locFunc <- mean
  } else if (loc == 'median'
    locFunc <- median
  }
    cat(sprintf "non such location metric %d", loc))
  }

  #if(ncol(data)>nrow(data)) data <- t(data)
  #data = abs(data)
  data_ = t(data)
  reference <- apply(data_,1,locFunc)
  # sometimes reference produces 0s so we turn them into 1s before division
  # so spectrum stays unchanged
  reference[reference==0] <- 1

  quotient <- data_/reference
  quotient.withLocFunc <- apply(quotient,2,locFunc)

  pqn.data <- t(data_)/quotient.withLocFunc
  pqn.data
}

# normalisation to total integral
TotArea <- function(data) {
  data_ = t(data)
  meanInt = sum(apply(data_,1,mean))
  scalingFactor = apply(data_,2,sum) / meanInt
  data_ = t(t(data_) / scalingFactor)
  t(data_)
}

# normalisation to reference peak
NormByBin <- function(data, bin) {
  data_ = t(data)
  refPeakInt = data[,bin]
  # adjust the integral for mean peak to preserve scale of spectra in the dataset
  refPeaksAdj = refPeakInt/mean(refPeakInt)
  data_ = t(t(data_)/refPeaksAdj)
  t(data_)
}
```

Appendix 3-8. Relative abundance box plots of mannose-related metabolites for P. graminicola using CRODAFAT and Olive oil as FA feedstock over a 96 hours time course. N= 6, $p < 0.05$.



Appendix. 4-1. Detailed list of NCBI accession number for MEL cluster proteins from different Basidiomycetes.

### EMT1 List

CDU26159.1 *Sporisorium scitamineum*
CBQ73522.1 *Sporisorium reilianum SRZ2*
CDI53946.1 *Melanopsichium pennsylvanicum 4*
XP_012190145.1 *Pseudozyma hubeiensis SY62*
XP_011389468.1 *Ustilago maydis 521*
CCF52717.1 *Ustilago hordei*
SAM82152.1 *Ustilago bromivora*
XP_014653801.1 *Pseudozyma antarctica*
ETS61959.1 *Pseudozyma aphidis DSM 70725*
GAC75887.1 *Pseudozyma antarctica T-34*
XP_681404.1 *Aspergillus nidulans FGSC A4*

### MAC1 List

CDU26160.1 *Sporisorium scitamineum*
CBQ73521.1 *Sporisorium reilianum*
XP_012190147.1 *Pseudozyma hubeiensis SY62*
XP_011389467.1 *Ustilago maydis 521*
CDI53947.1 *Melanopsichium pennsylvanicum 4*
CCF52716.1 *Ustilago hordei*
GAC75889.1 *Pseudozyma antarctica T-34*
ETS61961.1 *Pseudozyma aphidis DSM 70725*
XP_014653798.1 *Pseudozyma antarctica*
XP_681406.1 *Aspergillus nidulans FGSC A4*

### MAC2 List

CDU26158.1 *Sporisorium scitamineum*
CCF52718.1 *Ustilago hordei*
SAM82151.1 *Ustilago bromivora*
CDI53945.1 *Melanopsichium pennsylvanicum 4*
XP_011389530.1 *Ustilago maydis 521*
XP_012190144.1 *Pseudozyma hubeiensis SY62*
ETS61960.1 *Pseudozyma aphidis DSM 70725*
XP_014653799.1 *Pseudozyma antarctica*
GAC75888.1 *Pseudozyma antarctica T-34*
CBQ70845.1 *Sporisorium reilianum SRZ2*
XP_011387307.1 *Ustilago maydis 521*
CDS00082.1 *Sporisorium scitamineum*
CDR88152.1 *Sporisorium scitamineum*
XP_681405.1 *Aspergillus nidulans FGSC*

### MAT1 List

CBQ73519.1 *Sporisorium reilianum SRZ2*
CDU26162.1 *Sporisorium scitamineum*
XP_011389465.1 *Ustilago maydis 521*
XP_012190149.1 *Pseudozyma hubeiensis SY62*
GAC96562.1 *Pseudozyma hubeiensis*

SAM82157.1 *Ustilago bromivora*
CCF52714.1 *Ustilago hordei*
ETS61963.1 *Pseudozyma aphidis DSM 70725*
CDI53949.1 *Melanopsichium pennsylvanicum 4*
GAC75891.1 *Pseudozyma antarcticus T-34*
XP_011387305.1 *Ustilago maydis 521*
CDS00081.1 *Sporisorium scitamineum*
XP_681403.1 *Aspergillus nidulans FGSC A4*

**MMF1 List**

CBQ73520.1 *Sporisorium reilianum SRZ22*
XP_012190148.1 *Pseudozyma hubeiensis SY62*
XP_011389466.1 *Ustilago maydis 521*
CCF52715.1 *Ustilago hordei*
CDI53948.1 *Melanopsichium pennsylvanicum 4*
SAM82156.1 *Ustilago bromivora*
ETS61962.1 *Pseudozyma aphidis DSM 70725*
XP_014653797.1 *Pseudozyma antarctica*
GAC75890.1 *Pseudzoyma antarctica T-34*
CBQ72959.1 *Sporisorium reilianum SRZ2*

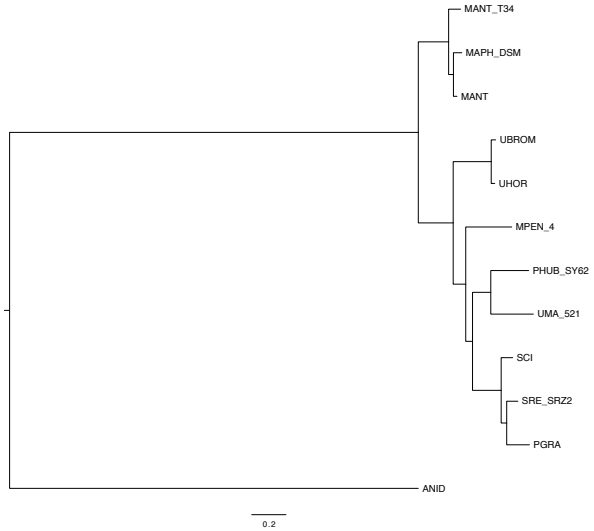
Appendix. 4-2. Example of HTSeq-count code used in this study.

```
htseq-count -f bam -t CDS -i gene_id -m union ../F3_1.1/accepted_hits.bam
/pub39/ext/stefany/braker/all_libraries/augustus_CDS.gff > F3_1.1_counts

#change the file by using the CDS as feature and gene_id
#change by using stranded yes as default setting
```

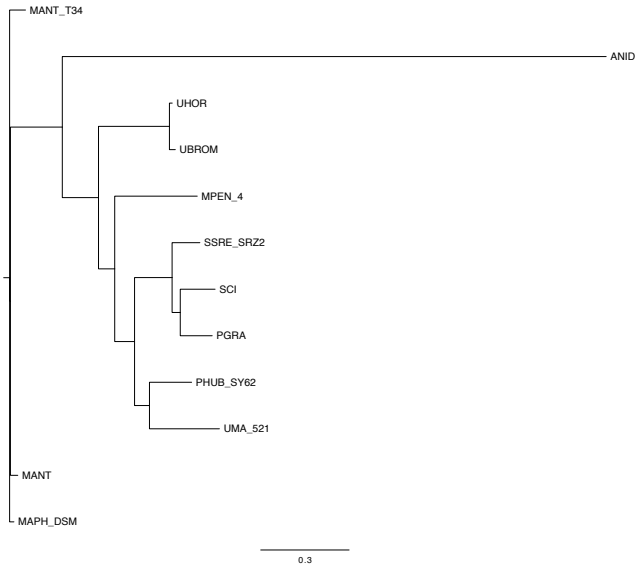
Appendix. 4-3. Primer list for qRT-PCR analysis used in this study

| Primer | Sequence 5'-3' |
|---|---|
| F_mac2_qRT | ggtcctcaggaacaaagacg |
| R_mac2_qRT | tgaatgtctcgcacgttctc |
| F_mac1_qRT | CCATCCTAGCTCGTCACACA |
| R_mac1_qRT | TGGTCGCAGGTATCGTGTT |
| F_mmf1_qRT | GACCAGGCTTGCAGTTCTTC |
| R_mmf1_qRT | TACTGGCCAGCAGTGTCAAC |
| F_mat1_qRT | CTGCAGCTGATCCTGGAAAC |
| R_mat1_qRT | CTTGTTGCGCGCCTCTTT |
| F.Actin_qRT | GTGCGCTTCTGTACAGCTTG |
| R.Actin_qRT | GACGCTCTCCTTGAAGTCGT |
| F_emt_qRT | CTCGAAATCGAGCCTGACAT |
| R_emt_qRT | ATGGTGAGCAAGGCACTGT |

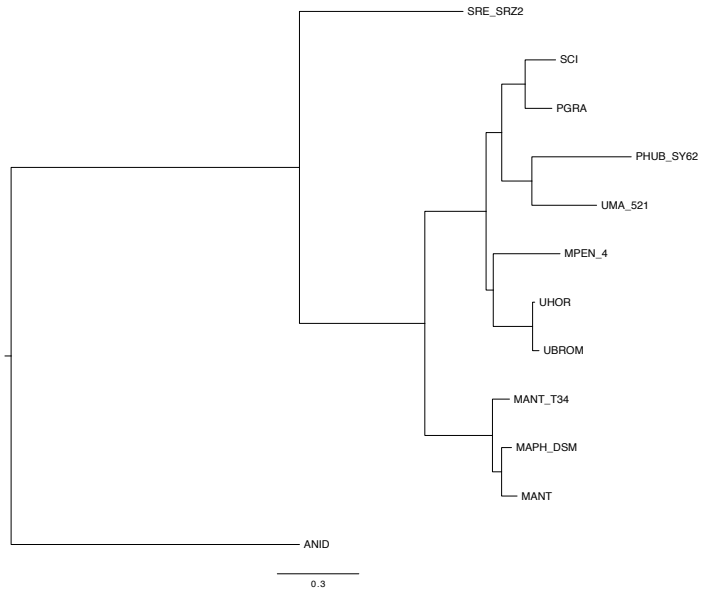
Appendix. 4-4. Molecular phylogenetic tree constructed using the amino acid sequence of EMT1 for P. graminicola and other related fungi. Species key: UBROM: Ustilago bominivora, UHOR: Ustilago hordei, MPEN_4: Melanopsichium pennsylvanicum 4, UMA_521: Ustilago maydis, PHUB_SY62: Pseudozyma hubeiensis SY 62, PGRA: Pseudozyma graminicola, SRE_SRZ2: Sporisorium reilianum SRZ2, SSCI: Sporisorium scitamineum, MANT_T34: Moesziomyces antarcticus T34, MANT: Moesziomyes antarcticus, MAPH_DSM70725: Moesziomyces aphidis DSM 70725, ANID: Aspergillus nidulans FGSC A4. The alignment was done using ClustalW, a rapid maximum-likelihood (ML) with a bootstrap of 100 runs was used. The tree was drawn using FigTree.




Appendix. 4-5. Molecular phylogenetic tree constructed using the amino acid sequence of MAC1 for P. graminicola and other related fungi. Species key: UHOR Ustilago hordei, MPEN_4: Melanopsichium pennsylvanicum 4, UMA_521: Ustilago maydis 521, PHUB_SY62: Pseudozyma hubeiensis SY62, PGRA: Pseudozyma graminicola, SRE_SRZ2: Sporisorium reilianum SRZ2, SSCI: Sporisorium scitamineum, MANT_T43: Moesziomyces antarcticus T34, MANT: Moesziomyces antacticus, MAPH_DSM70725: Moesziomyces aphidis DSM 70725, ANID: Aspergillus nidulans FGSC A4. The alignment was done using ClustalW, a rapid ML with a bootstrap of 100 runs was used. The three was drawn using FigTree.



Appendix. 4-6. Molecular phylogenetic tree constructed using the amino acid sequence of MAC2 for P. graminicola and other related fungi. Species key: UBROM: Ustilago bominivora, UHOR: Ustilago hordei, MPEN_4: Melanopsichium pennsylvanicum 4, UMA_521: Ustilago maydis 521, PHUB_SY62: Pseudozyma hubeiensis SY62, PGRA: Pseudozyma graminicola, SRE_SRZ2: Sporisorium reilianum SRZ2, SSCI: Sporisorium scitamineum, MANT_T34: Moesziomyces antarcticus T34, MANT: Moesziomyces antarcticus, MAPH_DSM70725: Moesziomyces aphidis DSM 70725, ANID: Aspergillus nidulans FGSC A4. The alignment was done using ClustalW. A rapid ML with a bootstrap of 100 runs was used. The tree was drawn using FigTree.

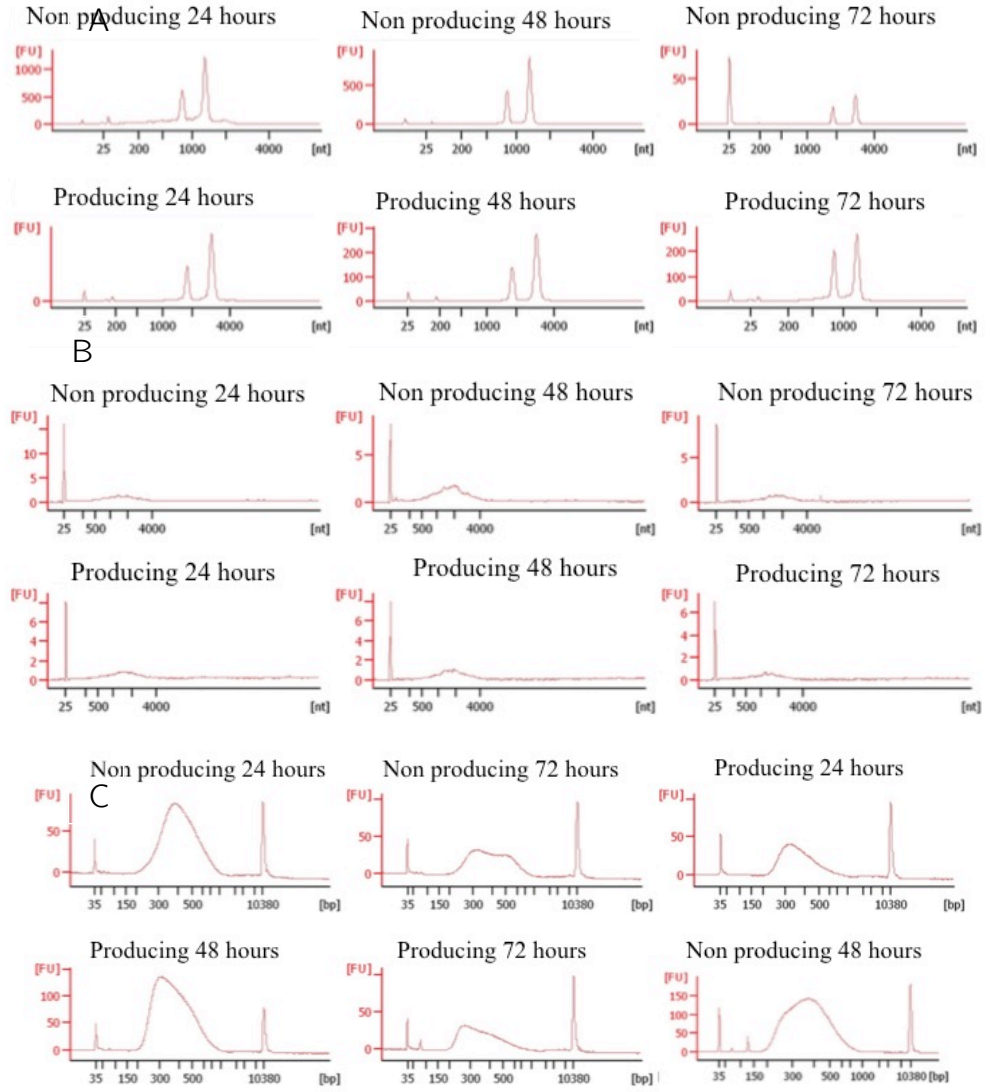



Appendix. 4-7. Qubit values for RNA and cDNA concentration.

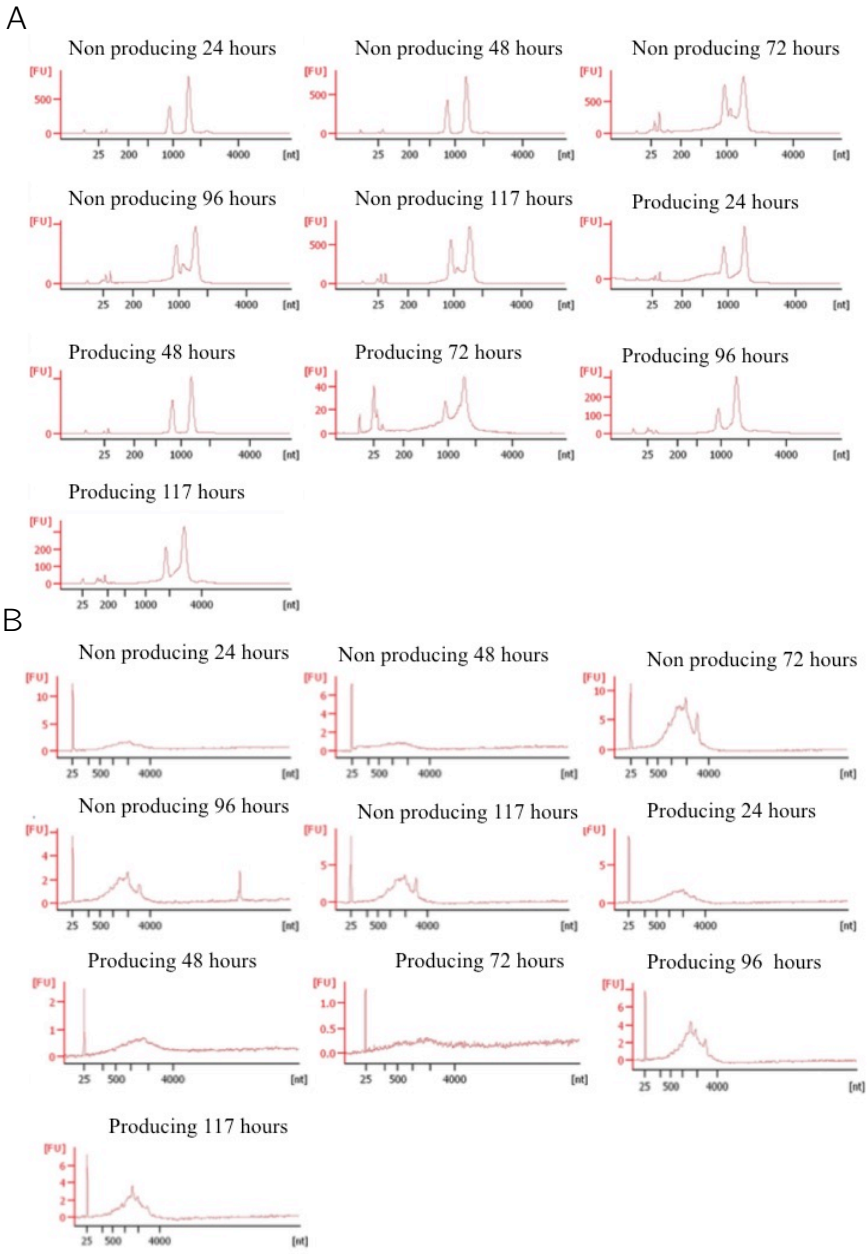
| Time point (hours) | Condition | Total RNA (ng/ml) | mRNA (ng/ml) | cDNA (ng/ml) | Fermentation |
|---|---|---|---|---|---|
| 24 | NP | 1200 | 0.91 | 0.50 | First |
| 48 | NP | 1900 | 0.149 | 6 | First |
| 72 | NP | 200 | 0.734 | 4.5 | First |
| 96 | NP | 160 | 0.219 | 4.5 | First |
| 117 | NP | 180 | 0.319 | 24 | First |
| 24 | NP | 99 | 0.2 | 1.40 | Second |
| 48 | NP | 84 | 0.281 | 2.8 | Second |
| 72 | NP | 44.3 | 0.108 | 0.5 | Second |
| 24 | P | 160 | 0.153 | 0.5 | First |
| 48 | P | 40 | 0.069 | 0.60 | First |
| 72 | P | 57 | Out of range | 4.5 | First |
| 96 | P | 125 | 0.318 | 5 | First |
| 117 | P | 130 | 0.244 | 9.5 | First |
| 24 | P | 44.1 | 0.125 | 4.73 | Second |
| 48 | P | 76 | 0.121 | 2.29 | Second |
| 72 | P | 17.3 | 0.100 | 0.734 | Second |

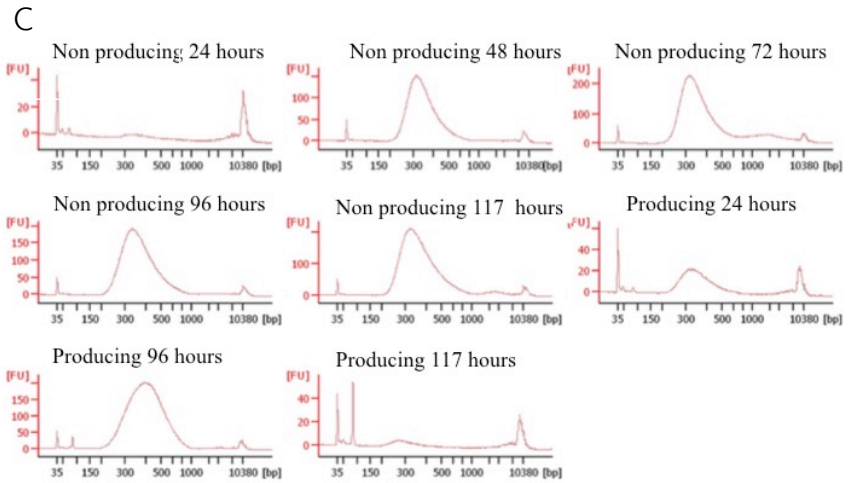
NP= non-producing conditions, P= producing conditions.

Appendix. 4-8. Quality control and integrity for RNA-seq libraries. Samples analysed on the bioanalyzer from induced and non-induced conditions at 24, 48 and 72 hours from a fermenter system. Induced condition refers to the presence of both, glucose and FA as carbon source on the media. Non-induced condition refers to the presence of only glucose as carbon source on the media. A) Total RNA, B) mRNA after depletion, C) cDNA libraries. The Y-axis represents Fluorescence Units (FU) and the X-axis represents base pairs length. The peaks shows the size of the sequenced fragments.

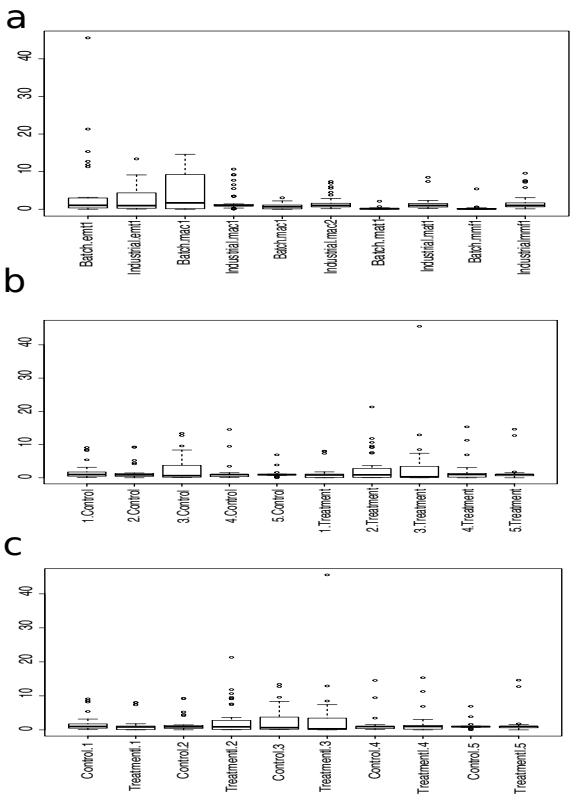


Appendix. 4-9. Quality control and integrity for RNA-seq libraries. Samples analysed on a bioanalyzer for induced and non-induced conditions at 24, 48, 72, 96 and hours from a fermenter system. Induced condition refers to the presence of both, glucose and FA as carbon source on the media. Non-induced condition refers to the presence of only glucose as carbon source on the media. A) Total RNA, B)mRNA after depletion, C)cDNA libraries. Induced conditions for 48 and 72 hours not shown. The Y-axis represents Fluorescence Units (FU) and the X-axis represents base pairs length. The peaks show the size of the sequenced fragments.






Appendix. 4-10. qRT-PCR boxplots for combination of evaluated variables from experimental design. Boxplots showing the mean, first and third quartiles, and maximum and minimum values for interaction between tested variables. Results shown are for biological replicates on n = 3 (Batch system) or n= 3 (fermenter system) technical replicates. a) Shows interaction between variables source and gene. b) Data distribution for interaction between time and condition. c) Data distribution for interaction between condition and time.



Appendix. 4-11. Table 2. Candidate models with effects of the covariates on α and ϕ and Likelihood Ratio test (LR) showing the best model for MEL gene expression. The candidate model selected is highlighted in bold. Firstly, each covariate was individually compared with the null model, then, the best model were selected and compared with the addition of covariates and so on. ANOVA with a Chi-Square test values reported to identify the significance of the variable and its iteration in terms of expression levels ($p<0.05$).

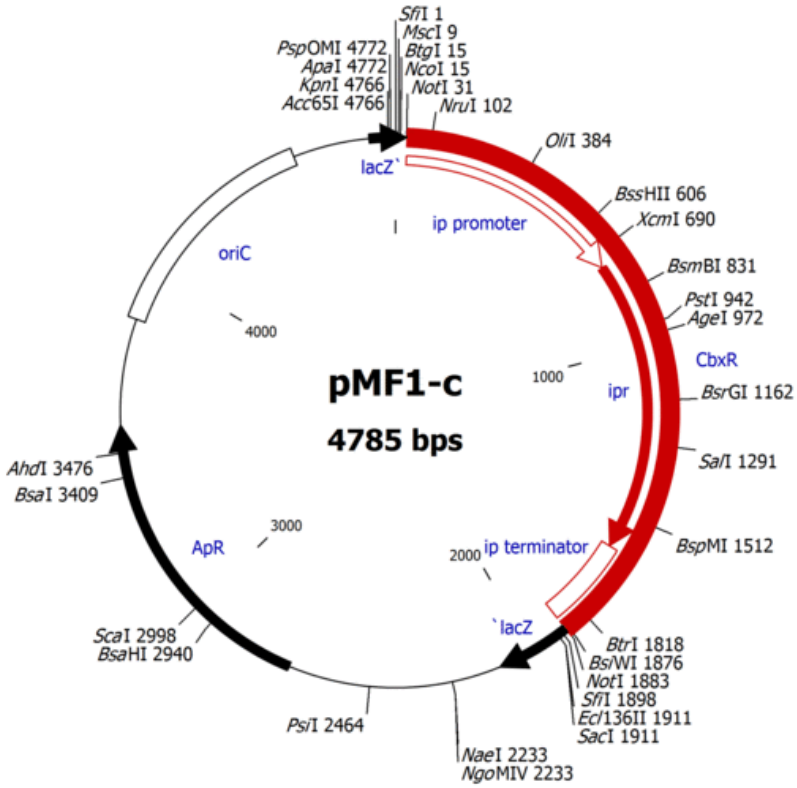
| | Log Likelihood | k* | df** | LR |
|---|---|---|---|---|
| Null model | -1086.418 | 0 | 1 | |
| Time | -567.4297 | 1 | 3 | 2.20E-16 |
| Condition | -566.2295 | 1 | 3 | 2.20E-16 |
| Source | -567.0177 | 1 | 3 | 2.20E-16 |
| **Gene** | -538.8295 | 1 | 6 | 2.20E-16 |
| Gene+Source | -538.3513 | 2 | 7 | 3.80E-01 |
| Gene+Time | -538.7063 | 2 | 7 | 0.6626 |
| Gene+Condition | -537.6645 | 2 | 7 | 0.1645 |
| Gene*Time | -535.6803 | 3 | 11 | 0.4354 |
| Gene*Condition | -534.3939 | 3 | 11 | 0.1707 |
| **Gene*Source** | -503.4828 | 3 | 11 | 2.36E-09 |
| Gene*Source+Time | -503.3989 | 4 | 12 | 7.42E-01 |
| Gene*Source+Condition | -502.3351 | 4 | 12 | 3.08E-02 |
| Gene*Source*Time | -502.1412 | 8 | 14 | 0.000165 |
| **Gene*Source*Condition** | -478.1268 | 8 | 21 | 3.08E-06 |

* Number of parameters

**Degrees of freedom are based in the difference between the numbers of parameters of each pair of comparison.

Appendix 5-1. Carboxin resistance gene plasmid from *Ustilago* Community

(http://www.mikrobiologie.hhu.de/ustilago-community.html)




Appendix 5-2. List of primers used to create deletion cassette

| Primer name | Sequence 5'-3' |
|---|---|
| UM CASSETTES R | TAAAACGACGGCCAGTGAAT |
| UM CASSETTES F | ACCATGATTACGCCAAGCTC |
| PG EMT1 F1 | TACCATTGCCTCCTTGTCCT |
| PG EMT1R4 | TGCTTGTTGAACCAGGATGA |
| PG EMT1 GIBSON F | ACTGGCCGTCGTTTTATCAGTCAGTCGTTCCGCC |
| PG EMT1 GIBSON R | CTTGGCGTAATCATGGTTGCTTGTTGAACCAGGATG |
| PG EMT1 F2 | TCTCGCAACTGATTGTCCAG |
| PG EMT1R3 | ATCACTGCGATCACAACAGG |
| EMT1_NEWF1 | ATCTGCTCGCTTGAAGATGG |
| EMT1_NEWR4 | TTCCGAGTTATGCTTGTACCG |
| EMT1_F2 | TTGTTGTTGCCGTAGGACAC |
| EMT1_R3 | ATTACCAAACACGGCAGGAG |

Appendix 5-3. List of primers used for diagnostic PCR for mutant confirmation

| Primer name | Sequence 5'-3' |
|---|---|
| F. AREA | GACCGACTTTGACAGCTTGC |
| R.AREA | GCTGCTCGATGGACTGTATG |
| F. GTI | CGTCTCGAGGACCCATCTTA |
| R. GTI | GAGTTGAACGCCTTCGTGTT |
| F.PAC1 | TCTCGCTCTCAAACCATGC |
| R.PAC1 | TGATGAGATTGTCGGTGGTG |
| F. CARBOXIN | GTGCCAGACTTGACCCAGTT |
| R.CARBOXIN | TCGGTAGAGCGAAAAGGTGT |
| F.EMT1 | ATGTTCCTGTGGACCGACTC |
| R.EMT1 | CAAGATCCGGTCTCTTCTCG |

Appendix 5-4. Relative abundance box plots for normalised (chloroform peak) MEL related metabolites over a 96 hours time course micro fermentation for *P. graminicola* WT and ΔPgAREA-1. N=6. Numbers corresponds to sampling time points in hours. Addition of FA denoted by "FA" before the time point stamp.